



**EXPLORING TEACHERS' DISPOSITIONS TOWARD AI FEEDBACK
ACCURACY AND THE REFINEMENT PROCESS AND THEIR
ENHANCEMENT PRACTICES IN EFL WRITING**

Doktora Tezi

Burcu GÜNDOĞDU

Eskişehir 2025

**EXPLORING TEACHERS' DISPOSITIONS TOWARD AI FEEDBACK
ACCURACY AND THE REFINEMENT PROCESS AND THEIR
ENHANCEMENT PRACTICES IN EFL WRITING**

Burcu GÜNDOĞDU

PhD DISSERTATION

Program in English Language Teaching

Department of Foreign Language Education

Supervisor: Prof. Dr. Ali MERÇ

Anadolu University

Graduate School

June 2025

Eskişehir

JÜRİ VE ENSTİTÜ ONAYI

Burcu GÜNDOĞDU'nun "Öğretmenlerin Yabancı Dil Olarak İngilizce Yazmada Yapay Zekâ Geri Bildirim Doğruluğuna ve İyileştirme Sürecine Yönelik Görüşlerinin ve Geliştirme Uygulamalarının İncelenmesi" başlıklı tezi .../.../20.. tarihinde aşağıdaki jüri tarafından değerlendirilerek "Anadolu Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği"nin ilgili maddeleri uyarınca, Yabancı Diller Eğitimi Anabilim Dalı İngilizce Eğitimi Programında, Doktora tezi olarak kabul edilmiştir.

	<u>Unvan Adı Soyadı</u>	<u>İmza</u>
Üye (Danışman)		
Üye		
Üye		
Üye		
Üye		

.....
Lisansüstü Eğitim Enstitüsü Müdürü

FINAL APPROVAL FOR THESIS

This thesis titled “Exploring Teachers’ Dispositions toward AI Feedback Accuracy and The Refinement Process And Their Enhancement Practices in EFL Writing” has been prepared and submitted by Burcu GÜNDÖĞDU in partial fulfillment of the requirements in “Anadolu University Directive on Graduate Education and Examination” for the Doctor of Philosophy (PhD)/Proficiency in Arts in Department of Foreign Language Education has been examined and approved on/...../.....

Committee Members

Signature

Member (Supervisor)

.....

.....

Member

.....

.....

Member

.....

.....

Member

.....

.....

Member

.....

.....

.....

Director of Graduate School

ÖZET

ÖĞRETMENLERİN YABANCI DİL OLARAK İNGİLİZCE YAZMADA YAPAY ZEKÂ GERİ BİLDİRİM DOĞRULUĞUNA VE İYİLEŞTİRME SÜRECİNE YÖNELİK GÖRÜŞLERİNİN VE GELİŞTİRME UYGULAMALARININ İNCELENMESİ

Burcu GÜNDOĞDU

Yabancı Diller Eğitimi Anabilim Dalı

İngilizce Eğitimi Doktora Programı

Anadolu Üniversitesi, Lisansüstü Eğitim Enstitüsü, Haziran 2025

Danışman: Prof. Dr. Ali MERÇ

Bu çoklu durum çalışması, İngilizce hazırlık sınıfı öğretmenlerinin, yapay zekâ tarafından üretilen geri bildirim teknik ve pedagojik açıdan öğretmen benzeri hâle getirmek için nasıl iyileştirdiklerini; hata kodlamalı ve içerik odaklı geri bildirimin doğruluğunu kendi iyileştirme uygulamaları doğrultusunda nasıl değerlendirdiklerini; ayrıca bu iyileştirme sürecini nasıl değerlendirdiklerini incelemektedir. Önceki araştırmalar sınıf içi kullanım öncesi öğretmen filtresinin önemini vurguladığından, öğretmenlerin renk kodlu ve yapılandırılmış incelemelerle belgelenen düzenlemeleri birincil veri kaynağı olarak kullanılmıştır. Bu düzenlemelerden hareketle, öğretmenler yapay zekâ tarafından oluşturulan geri bildirimin doğruluğunu değerlendirerek gerekli müdahale düzeyini belirlemiş, mülakatlar ise sınıf içi uygulanabilirliğini artırmaya yönelik öneriler sunmuştur. Analiz, hata kodlu geri bildirimin anlam, açıklık ve yapı gibi unsurları göz ardı ederek dilbilgisine fazla odaklandığını ortaya koymuştur. Bu sınırlılıklar, öğretmenlerin hata kodlu geri bildirim olan güvenini azaltmış ve süreci onlar için daha zahmetli hâle getirmiştir. İçerik geri bildirim ise paragraf düzeyinde bütüncül okuma eksikliğinden dolayı tekrarlayan fikirleri sıklıkla gözden kaçırmıştır. Buna karşılık, öğretmenler içerik geri bildirimini daha kullanışlı bulmuş, ancak öğrencilerin bireysel ihtiyaçlarına göre öncelik alanları eklediklerinde tamamen kullanışlı bulabileceklerini dile getirmişlerdir. Bu bulgular, yapay zekâ tarafından oluşturulan geri bildirimin doğruluğunu ve sınıf içi uygulanabilirliğini artırmaya yönelik gelecekteki çalışmalara ışık tutabilir.

Anahtar kelimeler: Yapay Zekâ tabanlı yazma geri Bildirimi, İngilizce Öğretmenlerinin iyileştirme uygulamaları, Geliştirme önerileri, Doğruluk değerlendirmeleri

ABSTRACT

EXPLORING TEACHERS' DISPOSITIONS TOWARD AI FEEDBACK ACCURACY AND THE REFINEMENT PROCESS AND THEIR ENHANCEMENT PRACTICES IN EFL WRITING

Burcu GÜNDOĞDU

Department of Foreign Language Education
PhD in English Language Teaching
Graduate School of Anadolu University, June 2025

Supervisor: Prof. Dr. Ali MERÇ

This multiple case study examines how English preparatory class instructors refine AI-generated feedback to render it more teacher-like in both technical and pedagogical terms, how they perceive the accuracy of AI-generated feedback for both error-coded and content-related on student essays based on their refinement insights, and how they evaluate the process of refining it. As prior studies emphasize the need for a teacher filter before classroom use, this study focuses on teachers' refinements, documented through color-coded and structured reviews, which served as the primary data sources for exploring this issue. Informed by their refinement practices, teachers evaluated the accuracy of AI-generated feedback and identified the extent of required teacher intervention, while the interview with teachers elicited practical suggestions to enhance the classroom usability of AI feedback. The analysis revealed a key limitation of AI-generated error-coded feedback: excessive focus on grammatical correctness leading AI to overlook meaning, clarity, and structural issues, and more. These limitations reduced teachers' trust in error-coded feedback and increased their reluctance to use it, as they found the process of refinement more burdensome. In contrast, content feedback often missed repetitive ideas due to its lack of focus on holistic, paragraph-level processing. However, teachers found content feedback more useful, noting it would be fully effective only when supplemented with priority areas specific to each student's needs. These findings can inform future efforts to enhance the accuracy and classroom applicability of AI-generated feedback.

Keywords: AI-generated writing feedback, EFL Teachers' refinement practices, Suggestions for improvement, Accuracy evaluations

.../.../2025

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Bu tezin bana ait, özgün bir çalışma olduğunu; çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarında bilimsel etik ilke ve kurallara uygun davrandığımı; bu çalışma kapsamında elde edilen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi; bu çalışmamın Anadolu Üniversitesi tarafından kullanılan “bilimsel intihal tespit programı”yla tarandığını ve hiçbir şekilde “intihal içermediğini” beyan ederim. Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçları kabul ettiğimi bildiririm.

Burcu GÜNDOĞDU

.../.../2025

STATEMENT OF COMPLIANCE WITH ETHICAL PRINCIPLES AND RULES

I hereby truthfully declare that this thesis is an original work prepared by me; that I have behaved in accordance with the scientific ethical principles and rules throughout the stages of preparation, data collection, analysis and presentation of my work; that I have cited the sources of all the data and information that could be obtained within the scope of this study, and included these sources in the references section; and that this study has been scanned for plagiarism with “scientific plagiarism detection program” used by Anadolu University, and that “it does not have any plagiarism” whatsoever. I also declare that, if a case contrary to my declaration is detected in my work at any time, I hereby express my consent to all the ethical and legal consequences that are involved.

Burcu GÜNDOĞDU

ACKNOWLEDGEMENTS

This thesis marks the culmination of a challenging yet deeply rewarding journey. Throughout the process, I have encountered many obstacles and moments of uncertainty, but also growth, perseverance, and invaluable support from those who were part of this demanding journey, sharing my feelings of excitement, anxiety, and panic, and standing by me through both good and difficult times.

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Prof. Dr. Ali MERÇ, for his continuous support, guidance, and encouragement throughout the course of this study. His constructive feedback and insightful suggestions have been instrumental in shaping this thesis into its final form. I am sincerely thankful for his patience, expertise, and belief in my work.

I would also like to express my sincere gratitude to the members of my dissertation monitoring committee, Prof. Dr. Özgür YILDIRIM and Asst. Prof. Dr. Ali Ulus KIMAV. Their thoughtful feedback, academic guidance, and continuous support not only enriched the quality of this study but also made a significant contribution to the development of this thesis and to my academic growth.

I would like to extend my sincere thanks to the esteemed jury members, Assoc. Prof. Dr. Mustafa CANER and Prof. Dr. Yusuf DEMİR, for their valuable time, insightful comments, and constructive feedback during my thesis defense. Furthermore, I would like to express my formal gratitude for TÜBİTAK 2211/A National Ph.D. Scholarship Program for their financial support during this process.

Special thanks to Asst. Prof. Dr. Merve ÖZER for her constant encouragement and belief in me, and to my colleagues and friends at Antalya Bilim University, Özge DEMİRBAĞ, Cenk TAKIM, Burcu ORUÇ, Birgen ZEYBEK, Ezgi UYAN and Buse ERDEN for their unwavering support, uplifting friendship, and insightful academic suggestions.

My heartfelt thanks go to the English instructors at Antalya Bilim University for their kind participation and valuable help during the data collection. I am sincerely grateful for their time, insights, and thoughtful contributions, which enriched this study in meaningful ways. Despite the demands of their teaching schedules and the heavy

workload they carry, their willingness to take part in this research enabled me to complete this thesis. Without their cooperation and assistance, it would not have been possible.

I would also like to express my deepest appreciation to my dearest and very first English teacher, Elif Merve OTSAR, whose influence has stayed with me over the years. During my university preparation years, she not only guided me academically but also sparked my passion for the English language. Thank you for believing in me before I believed in myself. You did not just teach me a language, you helped shape the person I am today, and for that, I will always be grateful.

I owe my deepest gratitude to my family, whose presence has been my greatest source of strength and motivation. Especially to my beloved mother, for her strength and endurance in holding me steady and motivated, even during the times when I was overwhelmed with extreme worry and pessimism. Whenever I struggled, she would say, “Accept it as something ordinary,” gently reminding me that what I was going through was just a natural part of the process. With those words, she always gave me strength and hope. To my father and brother, your constant “You can do it!” spirit carried me through the hardest moments. I hope this work makes you all proud. Also, to my dog Cesur, my loyal little shadow throughout this journey. You may have slept through most of it, but your quiet presence by my side, the gentle nudges during my most stressful days, and the comfort you gave without saying a word meant more to me than you will ever know. You reminded me to take breaks, to breathe, and to keep going. This thesis has a bit of your pawprint on every page. *Canım ailem, koşulsuz sevginiz, sabrınız ve bitmeyen desteğiniz için ne kadar teşekkür etsem az. Varlığınız, hayatımdaki en büyük şans. İyi ki varsınız.* This journey would not have been possible without you.

With deep gratitude and hope for the future, I close this chapter. It’s time to rest, reflect, and enjoy the moments ahead.

Burcu GÜNDÖĞDU

June 2025

.../.../...

ÜRETKEN YAPAY ZEKÂ KULLANIM BEYANI

Bu tezi hazırlarken Claude 3.5 Sonnet adlı üretken yapay zekâ programından destek aldığımı/almadığımı beyan ederim. Tezin hazırlığı aşamasında üretken yapay zekâ programlarından (Claude 3.5 Sonnet, GPT-4o ve Gemini-Advanced) geribildirim üretimi için model değerlendirmesi yapmak ve araştırma verisi (geribildirim çıktısı) üretmek amacıyla destek aldım. Üretken yapay zekâ programlarından aldığım bilgilerin doğruluğunu kontrol ettiğimi bildiririm.

Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçları kabul ettiğimi bildiririm.

Burcu GÜNDOĞDU

.../.../...

DECLARATION OF GENERATIVE ARTIFICIAL INTELLIGENCE USAGE

I declare that I received support from generative artificial intelligence programs (Claude 3.5. Sonnet) while preparing this thesis/dissertation. I have received support from generative artificial intelligence programs (Claude 3.5 Sonnet, GPT-4o and Gemini-Advanced) for the purposes of conducting model evaluation for feedback generation and producing research data (feedback outputs) during the preparation of the thesis/dissertation. I declare that I have checked the accuracy of the information I received from generative artificial intelligence programs.

I also declare that, if a case contrary to my declaration is detected in my work at any time, I hereby express my consent to all the ethical and legal consequences that are involved.

Burcu GÜNDOĞDU

TABLE OF CONTENTS

Page

EXPLORING TEACHERS' DISPOSITIONS TOWARD AI FEEDBACK ACCURACY AND THE REFINEMENT PROCESS AND THEIR ENHANCEMENT PRACTICES IN EFL WRITING	i
JÜRİ VE ENSTİTÜ ONAYI.....	ii
ÖZET	iii
ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ.....	v
STATEMENT OF COMPLIANCE WITH ETHICAL PRINCIPLES AND RULES	vi
ACKNOWLEDGEMENTS	vii
ÜRETKEN YAPAY ZEKÂ KULLANIM BEYANI	ix
DECLARATION OF GENERATIVE ARTIFICIAL INTELLIGENCE USAGE ..	x
TABLE OF CONTENTS	xi
LIST OF TABLES	xv
LIST OF FIGURES	xvii
LIST OF ABBREVIATIONS	xviii
1 INTRODUCTION	1
1.1 Background to the Study	2
1.2 Statement of the Problem	6
1.3 Purpose and Significance of the Problem	9
1.4 Limitations	13
1.5 Definition of Key Terms	14
2 REVIEW OF LITERATURE	18
2.1 A Historical Look at Feedback in L2 Writing	18
2.1.1 Behaviorist approach to Feedback (1950s)	19
2.1.2 Constructivist approach to feedback (1970s-1980s)	20
2.1.3 Cognitive process theory and feedback (1980s)	21
2.1.4 Social constructivist approach to feedback (1978, onwards)	21
2.1.5 Transition to machine feedback and generative AI chatbots (2010s- Present)	22
2.2 Conceptual & Theoretical Frameworks for AI feedback	23
2.2.1 Behaviorism and AI: is AI behaviorist?	24
2.2.2 Vygotsky's social constructivist theory: is AI the new MKO?	25
2.2.3 Krashen's comprehensible input: is AI feedback comprehensible?	26

2.2.4 Interaction hypothesis and AI as a conversational partner	27
2.2.5 Output hypothesis and contribution of AI to noticing and self-correction	27
2.2.6 AI as a tool for formative feedback.....	28
2.3 Overview of Theories that Non-supporting AI Feedback.....	28
2.3.1 Noam Chomsky's innateness hypothesis	29
2.3.2 Searle's Chinese Room Argument	30
2.4 Defining the Characteristics of Effective Feedback in L2 Writing.....	31
2.4.1 Broad definition and key features of effective feedback	31
2.4.2 Effective written corrective feedback (WCF) in literature.....	35
2.4.3 Teacher, peer, and AI WCF feedback: a comparative review of their characteristics in the literature.....	36
2.4.3.1 Teacher feedback	39
2.4.3.2 Peer feedback	40
2.4.3.3 AI feedback.....	42
2.5 Overview of Existing Empirical Studies on AI feedback	46
2.5.1 Studies on the Perceptions of AI-Generated Feedback.....	47
2.5.2 Comparative studies examining AI and human feedback on student writing	51
2.5.3 Comparisons among different gen-AI for feedback provision	55
2.5.4 Identified gaps and directions for future research	56
3 METHODOLOGY	58
3.1 Research Design	58
3.1.1 Multiple case study design	58
3.2 Context of the Study	59
3.3 Participants of the Study	62
3.3.1 Teacher profiles	64
3.4 Training	66
3.4.1 Training of AI platforms.....	66
3.4.2 Selection of Claude 3.5 sonnet as an AI for providing feedback.....	68
3.4.3 Expert Review of the AI Prompt and Data Collection Instruments	72
3.4.4 Training of Teachers	73
3.4.5 Refinements implemented after the pilot study	77
3.4.5.1 Adjustments for the color-coding review form	78
3.4.5.2 Adjustments for structured review table form	78

3.4.5.3 Adjustments for the Accuracy Evaluation (rubric).....	79
3.5 Data Collection Tools.....	79
3.5.1 Color-coded review form	79
3.5.2 Structured review table form	80
3.5.3 Accuracy Evaluations.....	80
3.5.4 Semi-structured Interview	81
3.6 Data Analysis.....	83
3.6.1 Overview of the data collection procedure.....	85
3.6.2 Researcher reflexivity.....	86
3.7 Trustworthiness of the Study.....	88
3.8 Ethical Considerations	90
4 FINDINGS.....	91
4.1 Findings related to Research Question 1	91
4.1.1 Quantitative analysis of teachers' refinement patterns for error-coded feedback.....	91
4.1.2 Analysis of teachers' rationales for the refinements of error-coded feedback.....	95
4.1.3. Qualitative analysis of teacher refinements for AI-generated error-coded feedback.....	96
4.1.3.1. Overlooked rewrite (RW) code.....	97
4.1.3.2. Overlooked codes: failure to detect meaning-level problems	101
4.1.3.3. Inaccurate codes: failure to provide appropriate codes for correction.....	101
4.1.3.4 Redundant codes: unnecessary error marking.....	103
4.2 Analysis of Teachers' Refinement for AI-generated Content Feedback	104
4.2.1 Quantitative analysis of teachers' content feedback refinement patterns	104
4.2.2 Qualitative analysis of teachers' content feedback refinements.....	105
4.2.2.1. Unaddressed Repetitive Ideas (CDF Category)	107
4.2.2.2 Missed or insufficient feedback for body paragraph development (CDF category).....	109
4.2.2.3 Missing structural feedback (CDF category)	111
4.2.2.4. Direct feedback (DF)	113
4.2.2.5. Feedback tone (FT).....	114
4.2.2.6. Missing feedback – error (MFE)	115
4.2.2.7. Missing feedback – language use (MF-LU)	116

4.2.2.8. Positive feedback (PF)	117
4.3. Findings related to Research Question 2	118
4.3.1 Quantitative findings related to AI-generated feedback accuracy for both error-coded and content-related feedback	118
4.3.1.1 Teacher acceptance levels for content feedback	121
4.3.1.2 Analysis of rubric evaluations and comments of teachers on the accuracy of AI-generated feedback.....	123
4.3.2 Qualitative findings related to AI-generated feedback accuracy from the teacher interviews	125
4.3.2.1. Qualitative findings for error-coded feedback accuracy	125
4.3.2.2. Qualitative findings for content feedback accuracy	130
4.4 Findings related to Research Question 3	132
4.4.1 Challenges of refining AI-generated feedback from teachers’ perspectives	133
4.4.2 Positive aspects of refining AI-generated feedback from teachers’ perspectives	135
4.4.3 Teachers’ attitudes for future usability of AI-generated feedback.....	137
4.4.4 Teachers' improvement suggestions for AI-generated feedback	139
5 CONCLUSION, DISCUSSION, IMPLICATIONS AND SUGGESTIONS FOR FURTHER STUDIES.....	142
5.1 Conclusion	142
5.2 Discussion	148
5.2.1 Teachers’ Refinement of AI Feedback: Patterns and pedagogical insights	148
5.2.2 Teachers’ evaluations of the accuracy of AI-generated feedback.....	152
5.2.3 Challenges and advantages of refining AI-generated feedback	153
5.2.4 Future classroom usability of AI-generated feedback	156
5.3 Implications and Suggestions for Future Studies	157
5.3.1 Pedagogical implications	158
5.3.2 Design and training implications for AI tools	161
5.3.3 Implications for future research.....	163
5.3.4 Implications and Practical Guidelines for Teachers to Train AI.....	165
REFERENCES	169
APPENDICES	
ÖZGEÇMİŞ	

LIST OF TABLES

	<u>Page</u>
Table 2.1 Key Theoretical Approaches and Their Contributions to Feedback Practices	18
Table 2.2 Theoretical Frameworks Supporting AI Feedback and Their Alignment with Present Literature Findings	24
Table 2.3 Theoretical Frameworks Supporting AI Feedback and Their Alignment with Present Literature Findings	29
Table 2.4 Supporting Literature on the Key Characteristics of Effective Feedback.....	32
Table 2.5 Comparison of Feedback Characteristics Across Teacher, Peer, and AI Feedback.....	37
Table 2.6 AI Feedback Limitations and Possible Teacher Complementation's for AI Feedback.....	46
Table 2.7 Overview of Key Empirical Studies on AI Feedback and the Present Study's Contribution	47
Table 3.1 Demographic Information about the Participant-Teachers	63
Table 3.2 A summary of Technical Capabilities of AI models.....	69
Table 3.3 A Summary of the Comparative Analysis of Feedback Performances Across AI Models.....	70
Table 3.4 A Summary of the Comparative Readability Analysis of AI Models.....	71
Table 3.5 Summary of Data Collection Tools, Descriptions of tools, and Contribution to Research Questions.....	82
Table 3.6 Overview of the Data Collection Procedure	86
Table 4.1 Distribution of Teachers' Refinements for Error-Code Feedback Across All Papers	92
Table 4.2 Refinement Distribution across Teachers	93
Table 4.3 Teachers' Refinement Distribution by Paper.....	94
Table 4.4 Acceptance Levels and Rationale Code Analysis for Content Feedback	105

Table 4.5 Summary of Content Feedback Issues and Strengths	106
Table 4.6 Unaddressed Repetitive Ideas	108
Table 4.7 Missed or Insufficient Feedback for Body Paragraph Development	109
Table 4.8 Missed or Insufficient Feedback for Body Paragraph Development	111
Table 4.9 Direct Feedback (DF).....	113
Table 4.10 Feedback Tone (FT).....	114
Table 4.11 Missing Feedback – Error (MFE)	115
Table 4.12 Missing Feedback – Language Use (MF-LU)	116
Table 4.13 Positive Feedback (PF)	117
Table 4.14 Acceptance Levels Across Teachers for Error Feedback	119
Table 4.15 Acceptance Levels Across Teachers for Error-coded Feedback	119
Table 4.16 Acceptance Levels Across Papers for Error-coded Feedback	120
Table 4.17 A General Overview of Teachers’ Acceptance Levels for Content Feedback.....	121
Table 4.18 Acceptance Levels for Content Feedback Across Teachers	122
Table 4.19 Acceptance Levels by Paper for Content Feedback.....	123
Table 4.20 Rubric Ratings Across Papers.....	124
Table 4.21 Error Feedback Accuracy Issues.....	126
Table 4.22 Content Feedback Gaps	131
Table 4.23 Pros and Cons of Refining AI feedback.....	133
Table 4.24 Teachers’ Attitudes for Future Usability of AI feedback	137
Table 4.25 Teachers' Improvement Suggestions for AI-generated Feedback	140

LIST OF FIGURES

	<u>Page</u>
Figure 1.1 Framework for AI literacy of Teachers, Ng et al. (2021).....	8
Figure 3.1 Overview of Research Design and data sources.....	59
Figure 3.2 Summary of the training steps for the Gen-AI platforms for feedback provision.....	68
Figure 3.3 Composition of the Expert Review Panel	73
Figure 3.4 Overview of Training Schedule.....	74
Figure 3.5 Workflow for Teacher Refinement process and task completion	75
Figure 3.6. The Summary of the Application of Summative and Conventional Content Analysis Methods to the Data.....	85
Figure 4.1 Frequency and Percentage of Rationale Use by Teacher for Error-Coded Feedback.....	95
Figure 5.1 Summary of Findings: Accuracy, Limitations, and Strengths of AI-Generated Feedback as Perceived by Teachers.....	143
Figure 5.2 Summary of the Pros and Cons of Refining AI-Generated Feedback, Usability of Feedback Based on Teachers’ Views, and Their Suggestions for Improvements	146
Figure 5.3 Key principles and teacher responsibilities in scaffolding Gen-AI for writing feedback.....	166

LIST OF ABBREVIATIONS

AI:	Artificial Intelligence
LLM:	Large Language Model
EFL:	English as a Foreign Language
CEFR:	Common European Framework of Reference for Languages
L2:	Second Language
MA:	Master of Arts
PhD:	Doctor of Philosophy
T:	Teacher (T1- T8 participants in the study)
Gen-AI:	Generative Artificial Intelligence
GPT-4o:	Generative Pre-trained Transformer by OpenAI, released in May 2024
Claude 3.5:	Claude 3.5 Sonnet by Anthropic, released in June 2024
Gemini-Advanced:	Gemini Advanced by Google DeepMind, released in May 2024

1 INTRODUCTION

As artificial intelligence (AI) has become increasingly integrated into education and classroom practices, both interest in its applications and concerns about its impact have grown. One such rapidly expanding area of implementation is writing instruction and feedback provision. In a relatively short time, AI-based writing feedback tools have evolved from simple proofreading applications that correct grammar and spelling errors to sophisticated Large Language Models (LLMs) capable of providing nuanced, contextual responses that approach the quality of human feedback. Consequently, the enhanced communicativeness and interactivity of these tools have accelerated their adoption as writing aids and feedback sources among both students and teachers.

Recent research demonstrates this growing reliance on Gen-AI Chatbots. Zhang and Yang (2024) found that students increasingly depend on ChatGPT for writing support, using it almost as frequently as they use search engine, Google. Similarly, AI has become an important resource for teachers across various tasks, including lesson planning, assessment design, supplementing teaching materials, and providing feedback (Nyaaba et al., 2024; Pettersson et al., 2024). This widespread adoption by teachers for multiple uses including feedback can be attributed to AI's ability to produce human-like responses that closely align with specific tasks and instructional goals when appropriately prompted by teachers (Baidoo-Anu & Owusu Ansah, 2023).

However, despite students' growing use of AI and the resulting increase in teachers' interest in employing it in their writing classes, many teachers remain hesitant to integrate AI-generated feedback into their classrooms. This hesitation stems primarily from documented shortcomings in AI-generated feedback identified by a growing body of research. Several studies highlighted AI Chatbots' insufficient contextual understanding and limited pedagogical relevance are major impediments to effective classroom implementation, necessitating teacher intervention (Algaraady & Mahyoob, 2023; Guo & Wang, 2023; Kurt & Kurt, 2024). These issues have raised ethical concerns, increasing doubts surrounding the accuracy of AI-generated feedback for educational use, and are compounded by teachers' limited understanding of how to implement AI feedback effectively and ethically in their classrooms.

All these concerns combined highlighted the need for teacher filter and monitoring for the successful integration of AI-generated feedback into teaching (Asadi et al., 2025; Kurt

& Kurt, 2024). Nevertheless, there is a notable lack of empirical research examining how teachers actually interact with and refine AI-generated feedback. This gap in the literature indicates lack of a clear understanding of the types and extent of refinement required to make AI feedback pedagogically sound and practically usable in real classroom contexts. It also remains unclear whether the effort required for such undertaking is worthwhile, given AI's current capabilities and limitations.

To address these critical gaps, this study adopts a qualitative approach to explore teachers' refinement practices of AI-generated feedback. Specifically, the study analyzes teachers' common refinement approaches, their acceptance patterns of AI-generated feedback, their editorial revision notes and suggestions, and the rationales underlying their editorial decisions. Based on the extent and nature of these refinements, the study also examines teachers' evaluations of AI-generated feedback accuracy. As a third aim, this research dwells on teachers' recommendations for improving AI-generated feedback systems. Collectively, these investigations aim to contribute empirically to our understanding of AI's current strengths in feedback provision and identify necessary improvements for AI systems to better reflect teacher pedagogical reasoning. The study therefore proposes several recommendations for training AI to generate pedagogically aware, teacher-like feedback, contributing to the growing research base that may investigate and replicate similar studies with updated AI versions appearing each year.

In different sections of this thesis, the research subject is explained in detail. Following this introductory overview providing research aim, the next section presents detailed background on generative AI use in education and its potential as an alternative feedback tool. Subsequent sections outline the statement of the problem, as well as the purpose and significance of the study, thereby helping readers understand the underlying motivation for conducting this research.

1.1 Background to the Study

Artificial intelligence has a long history of development and innovation. Early Automated Writing Correction and Feedback (AWCF) tools, were specifically designed to provide grammar-focused writing feedback based on large corpora. However, these tools allow only one-directional communication and lack the dynamic question-and-answer interaction that characterizes human feedback. On the other hand, generative AI chatbots, also known as Large Language Models (LLMs), operate through dialogue with users and

respond to their inquiries or prompts thanks to their communicative interface. This communicative evolution in AI has significantly altered teachers' traditional role as the sole provider of feedback in the eyes of students and has raised expectations among teachers for more advanced and pedagogically relevant AI support.

A substantial body of research acknowledges the potential of AI models to provide feedback and the combined power of teacher and AI feedback to enhance the effectiveness of teacher feedback. This recognition is grounded in earlier research on predecessor AI tools, which demonstrated that combining teacher feedback with AWCF tools can save teachers time by efficiently managing lower-order concerns (LOCs) in writing feedback while allowing teachers to focus more extensively on higher-order concerns (HOCs) (Jiang, Yu, & Wang, 2020; Ranalli & Yamashita, 2022).

Recent studies on Generative AI have similarly embraced its potential as a feedback provider. Zawacki-Richter et al. (2019) noted that, due to their advanced understanding of human language, AI is capable of providing instant, real-time, and personalized feedback on student writing, which can lighten the feedback workload of teachers caused by dealing with large volumes of student papers. Likewise, the study by Nazari, Shabbir, and Setiawan (2021) showed that AI tools positively facilitated writing instruction by enhancing students' writing abilities, promoting their autonomous and personalized learning as a learner, especially in higher education. McMurtrie (2022) acknowledged that AI can promote writing instruction by facilitating feedback provision for teachers, as long as ethical boundaries are well-drawn. As for students, thanks to its excellence in written communication and strong language skills, these tools can serve as a model by showcasing them how to compose clear and concise written output (Javaid et al., 2023).

Considering these benefits, the literature has explored whether AI can serve as a viable alternative or complementary tool to address documented shortcomings in teacher-provided feedback. Although extensive literature has consistently shown that teacher feedback has long been trusted for improving accuracy, expression, and organization in student writing (Chandler, 2003; Chen & Li, 2009; Ferris, 1997; Wang, 2006), and students have often deemed teacher feedback as more effective and motivating than other options (Fong & Schallert, 2023), several significant studies have also shown that the quality of teacher feedback can be adversely affected by teachers' cognitive load and time pressure.

Truscott (1999) highlighted a serious limitation of teacher feedback, suggesting that teachers' understanding of content-based problems in student papers can be influenced when they are required to give feedback on a large number of papers. Li, Kau, and Shiung (2023) elaborated on this finding, noting that teachers feel overwhelmed not only by the sheer volume of student papers but also mentally challenged when addressing both lower-order and higher-order feedback requirements for each paper. Higher-order feedback encompasses content elements such as logical organization, idea development with supporting details, coherence, and cohesion, while lower-order feedback focuses on writing mechanics including grammar, spelling, and punctuation.

Furthermore, time pressure and heavy workloads have been shown to negatively impact teacher feedback quality. Teachers may struggle to provide efficient, personalized, and timely feedback on student papers (Er, Dimitriadis, & Gašević, 2021). Additionally, Hyland and Hyland (2019) reported that time pressure can affect the language teachers use in feedback, leading them to abandon constructivist approaches in favor of overly critical language or comments.

Yet, despite AI's potential to complement teacher feedback, several factors continue to impede its widespread classroom integration. Initially, teachers' resistance to employing it due to the underlying fear that it might take over their jobs was a major barrier, yet this was largely compromised as teachers start to view AI as a supportive tool and assistant, rather than as a replacement for their professional expertise (Murphy, 2019).

However, several significant barriers remain unresolved. Mehdaoui (2024) identified multiple obstacles to integrating AI into classrooms, including technological constraints such as slow computers, poor internet connectivity, and overcrowded classrooms. He also highlighted teachers' concerns about potential student misuse and doubts regarding the accuracy and reliability of AI-generated feedback as key limitations to its widespread adoption. These challenges are further compounded by teachers' limited experience, insufficient training, and uncertainty about how to effectively incorporate AI tools into their teaching. Similarly, Baidoo-Anu and Owusu Ansah (2023) pointed to additional risks, including misinformation, biased AI outputs, and data privacy concerns, which continue to hinder broader implementation.

Another significant impediment is pedagogical challenge. As Vinge (1993) noted, there are different AI models designed for different general purposes, but as he noted, this leads

GEN-AI to remain limited at certain aspects easily performed by human, and hence far from achieving human-equivalent general intelligence. This limitation also means current AI systems cannot fully emulate teacher pedagogical reasoning.

Yet, what is promising is, AI technology continues advancing to address these challenges. Dwivedi et al. (2023) noted that recent AI platforms are released annually or more frequently, featuring diverse strengths and capabilities to meet various educational needs. Selecting the appropriate platform becomes crucial for achieving desired educational outcomes. Therefore, the present study utilized the pro-versions of three widely known Generative AI platforms available at the time of data collection: ChatGPT (GPT-4o, OpenAI), Claude 3.5 Sonnet (Anthropic), and Gemini Advanced (Google DeepMind). These three platforms demonstrate diverse characteristics and strengths based on their unique training focuses. ChatGPT-4o, developed by OpenAI (2024), excels in creative abilities and conversational skills, with creativity as its defining strength, though responses to identical prompts may vary between sessions. Claude AI, created by Anthropic (2024), is designed for high reasoning capacity, processing large data volumes while offering more consistent results with lower hallucination rates. Additionally, Claude emphasizes responsible AI usage by evaluating prompt ethics and demonstrating resistance to jailbreaks or user misuse. Unlike GPT and Claude, Gemini Advanced developed by (Google DeepMind, 2024) integrates with Google tools and accesses Google Search for real-time information, and this allows to generate up to date outputs.

Beyond platform selection, fine-tuning AI output is critical for generating classroom-ready feedback that closely resembles teacher-provided feedback. However, most existing studies have not attempted AI fine-tuning, representing a significant limitation in current pedagogical research. A comprehensive literature review reveals that most studies employed Generative AI in its pre-trained form to generate feedback without fine-tuning prior to use (Algaraady & Mahyoob, 2023; Guo & Wang, 2023; Guo, Pan, Li, & Lai, 2024; Steiss et al., 2024; Yoon, Miszoglud, & Pierce, 2023). These studies relied solely on prompting without uploading teacher-annotated samples to guide the model or engaging in iterative interactions to achieve desired outputs.

Consequently, it remains unclear how effectively these tools could perform when fine-tuned specifically for educational feedback purposes. AI platforms have not been sufficiently explored to determine whether they can be fine-tuned to match the accuracy,

constructive tone, and guiding nature characteristic of teacher feedback. Prior studies therefore fell short in evaluating whether AI can generate truly classroom-ready feedback that meets teacher expectations and demonstrates human-level pedagogical reasoning.

In response to this gap, the present study purposefully trained and fine-tuned the three aforementioned AI platforms to generate both error-coded and content-based metalinguistic feedback reflecting the feedback types typically provided by teachers in authentic classroom settings. During training, teacher-annotated samples and error code files with examples guided AI in producing such feedback. Drawing on teachers' refinement practices and pedagogical insights, it aims to provide guidelines for improving both technical and pedagogical accuracy of AI feedback and to inform future efforts for effective and ethical integration of AI in educational contexts.

1.2 Statement of the Problem

Although Generative Artificial Intelligence (GenAI) has opened new avenues for delivering feedback, its readiness for classroom implementation remains a concern. Challenges persist in terms of accuracy, ethics, and pedagogical alignment. However, amid such challenges, AI is developing rapidly, and often introducing advancements within a year or less. These include increased token limits that allow for greater input and output processing, enhanced language understanding, and improved contextual awareness. This ongoing evolution further reinforces that AI is not a passing trend but a rapidly evolving tool that necessitates constantly keeping up with the updates of this technology for both educators and students. Therefore, there is a growing need for research and ongoing evaluation to address the existing challenges concerning feedback accuracy and the ethical implementation of feedback in alignment with pedagogical principles.

The foremost concern before the employment of AI for feedback is its ethical use by students. The emergence of AI platforms have changed drastically altered students' conceptions of receiving feedback because they no longer consider teachers as the sole source (Shi, 2019). Being aware of AI tools are at their service, many students now also turn to AI platforms for feedback, as discussed in Henderson et al., (2025). However, this raises ethical considerations because, when used without teacher involvement, current AI platforms do not impose ethical restrictions or limitations on their output. This means that students may use them to answer all of their questions without putting in any effort

themselves. Therefore, Henderson et al. emphasized the importance of investigating effective ways to implement GenAI responsibly in higher education institutions. Otherwise, as Wang & Siau, (2019) discussed in their study, unregulated use of AI can facilitate AI-facilitated academic plagiarism, which is very difficult to identify reliably in students' writing, especially since current AI detectors can yield false positives.

Another related concern is students' overdependence on AI tools. Several studies in the literature have acknowledged that students tend to view AI as an easily accessible feedback source and sometimes use it to produce their assignment papers from scratch. Consequently, AI can pose a significant threat to students' critical thinking skills and their ability to apply independent reasoning when completing tasks (Aoun, 2017; Knox, 2020; Zawacki-Richter et al., 2019). Tseng and Warschauer (2023) also emphasized that AI writing tools, such as GPT, may diminish students' intellectual engagement with the task. Similarly, Koltovskaia (2020) found that students often accepted the output from Automated Writing Evaluation (AWE) tools as error-free and avoided double-checking it. This tendency may stem from their limited knowledge of how to approach and process AI-provided feedback with ethical awareness, as well as from a lack of understanding of what constitutes appropriate academic use (Wang, 2024). That is to say, they do not possess critical awareness of using AI in an ethically informed way to avoid cheating and plagiarism.

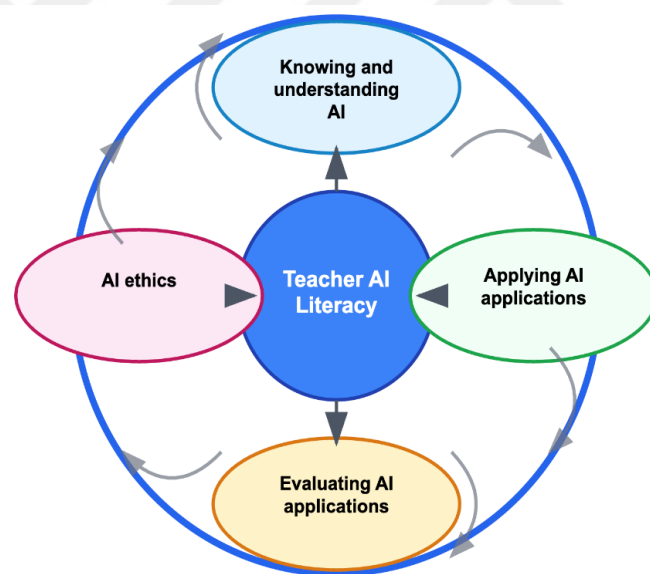
Beyond these ethical concerns, the literature also highlights that the readiness and accuracy of AI for classroom feedback remain questionable, as it may fail to fully comprehend the nuances of human language and the subtleties of emotion and pragmatics expressed in authentic communication. Due to this limitation, and because AI cannot handle excessive input, it is believed to produce some inappropriate or incomprehensible responses, a phenomenon known as "hallucination". This occurs when AI considers previous queries instead of focusing solely on the current data (Haleem, Javaid, & Singh, 2022; Lakshmanan, 2022). Additionally, AI can sometimes generate inaccurate feedback and excessive output, which further complicates the interpretation of AI feedback and can mislead students into making incorrect corrections (Pang, Kootsookos, & Cheng, 2024).

While these issues can be addressed through a teacher filter, a fundamental challenge is teachers' AI literacy. Many teachers have limited insight into the shortcomings of AI-generated feedback and the processes involved in integrating it into classroom practice.

Hence, the responsible integration of AI into educational practices remains an unresolved research question (Cavaleri et al., 2019; García-Peñalvo, 2023). Southworth et al. (2023) referred to this as a pedagogical challenge of the 21st century that still awaits resolution, particularly through teachers' involvement. In their view, in this era, AI literacy has become a fundamental requirement, and identifying the best possible practices for classroom implementation is critical, especially the teacher's role is more central than that of any other stakeholder in the educational sector. Similarly, Ng et al. (2021) also noted that expanding AI literacy can help overcome the obstacles to integrating AI effectively in classrooms. Building on this foundation, Zhao et al.(2022) employed Ng et al.'s framework to investigate teachers' AI literacy, and identified the following four key domains for ethical AI integration, which are presented in Figure 1.1 below:

Figure 1.1

Framework for AI literacy of Teachers, Ng et al. (2021)



Note. Adapted from Zhao et al.(2022), as cited in "Enhancing teacher AI literacy and integration through different types of cases in teacher professional development" by A.-C. E. Ding, L. Shi, H. Yang, & I. Choi, 2024, *Computers and Education Open*, 6, Article 100178, p. 2.

Based on this framework, developing teachers' AI literacy depends on four core components of teacher AI literacy. The first is knowing and understanding AI, which refers to having a clear grasp of AI concepts, its potential, and its limitations. The second is applying AI applications, which involves using AI tools effectively for teaching and learning purposes. The third is evaluating AI applications, which means being able to assess AI outputs for quality, reliability, and appropriateness in classroom contexts. The

fourth is AI ethics, which focuses on understanding and applying ethical principles when integrating AI into educational practices.

Building on the growing call to enhance teachers' AI literacy, several studies have further stressed the need for developing advanced-level GenAI competence among educators. For example, Miao and Cukurova (2024) stated that AI represents a new competency area, and all teachers must go beyond the basics. Kohnke (2024) also emphasized that rapid developments in AI make superficial knowledge insufficient; teachers must stay abreast of evolving features through continuous professional development. Building on this, Kukulska-Hulme et al. (2024) called for training that equips teachers to interact with AI effectively and critically evaluate its outputs from a professional standpoint. Collectively, these studies underscore that addressing the fundamental gap in teacher training and expanding their literacy is crucial for strengthening their capacity to effectively filter and refine AI-generated feedback.

This collaboration between AI capabilities and human expertise is vital for producing more accurate feedback (Algaraady & Mahyoob, 2023; Escalante, Pack, & Barrett, 2023). Gaining insight into teachers' acceptance levels of AI feedback, their correction practices, and the rationales behind implemented changes to AI feedback can reveal the pedagogical limitations of AI feedback and the areas where it requires improvements, and these can be used to generate feedback that is pedagogically similar to that of teachers.

The current study, with all these considerations in mind, aims to fill the existing gaps and pave the way for the ethically responsible implementation of AI-generated feedback in classrooms, taking into account teachers' views of AI-generated feedback with a focus on both technical and pedagogical accuracy.

1.3 Purpose and Significance of the Problem

There is a clear need for systematic investigation to fully understand the potential of AI in providing writing feedback. Existing studies vary in focus, feedback task specificity, and the findings, reporting different levels of accuracy and usefulness of AI-generated feedback. So, the black box of AI remains underexplored challenge (Zednik, 2021). This lack of transparency in AI's reasoning processes in feedback generation, combined with inconsistent results across studies, necessitates further testing across different feedback contexts.

When studies to date have been reviewed, it becomes evident that performance of AI in providing comprehensive error code feedback and targeted content feedback has been insufficiently explored. Moreover, research on how teachers can filter AI-generated feedback is notably absent, which highlights a significant gap in its potential use as classroom feedback.

To illustrate this gap, a recent study by Al-Garaady and Mahyoob (2023) examined GPT's performance in analyzing a short paragraph containing grammar and content errors. Their findings indicated that while GPT effectively identified surface-level errors, it failed to recognize all error types, particularly those requiring meaning-level and contextual understanding. This finding suggested that GPT's accuracy in error analysis diminishes when errors are related to meaning.

Similarly, Obaidoon and Wei (2024) investigated large language models' performance and tested out their ability to provide holistic written feedback on grammar, sentence structure, coherence, and writing quality in texts written by Chinese students. The results of this study favored Claude version 2 for its accurate and comprehensive grammar feedback. However, it was also noted that Claude version 2 was not flawless and could improve its contextualized explanations that could help guide student revisions more effectively. Unfortunately, aside from this one study, there is a dearth of comparative studies in the literature to verify Claude's performance in different types of feedback tasks.

Regardless of the feedback task, these studies consistently conclude that a teacher filter or a blended approach, combining the strengths of both teacher and AI feedback, is essential to address the limitations of AI-generated feedback (Escalante, Pack, & Barrett, 2023). Supporting this view, Guo and Wang (2023) found that collaboration between teachers and GPT can yield more personalized and inclusive feedback, yet noted the practical constraint that teachers may lack sufficient time to review AI outputs. Similarly, Obaidoon and Wei (2024) emphasized the need for continuous research to improve AI-generated feedback and highlighted the importance of teacher involvement in refining AI outputs to enhance quality. Steiss (2024) reinforced this need, reporting variety of accuracy issues in GPT-3.5 feedback. In the same vein, Cao and Zhong (2023) advised that AI should serve as a supplementary tool rather than a standalone feedback source. Echoing this recommendation, Barrot (2023) stressed that teacher refinement and

validation remain essential before AI feedback can be reliably integrated into classroom practice.

However, despite this growing recognition of the need for a teacher filter, there is still a dearth of research specifically investigating how teachers refine AI-generated feedback to improve its pedagogical accuracy in producing teacher-like feedback. Yet this is a gap that requires urgent attention, as it is becoming increasingly essential for educators to critically examine the capabilities of AI, make informed decisions based on this understanding, and utilize outputs that require minimal teacher intervention.

Another notable gap is that no study has yet explored the advantages and challenges that teachers encounter during the refinement process of AI-generated feedback. This means it remains unclear how teachers perceive refining AI-generated feedback, and whether they view the refinement process as worth undertaking. This is also a fundamental gap in the literature that should be addressed, as understanding how teachers perceive and engage with AI-generated feedback during the refinement process is essential for exploring effective ways to filter and enhance AI feedback for better outcomes.

To address these gaps in the literature, this study undertakes a comprehensive investigation into teachers' refinement reviews of AI-generated feedback, including not only examination of the types of refinements to AI-generated feedback, but also how they evaluated each feedback item by selecting acceptance levels, writing editorial notes, and providing detailed rationales to justify their refinements and enhancements. This multi-layered review process offers a transparent and in-depth account of teachers' decision-making and their pedagogical judgment in refining AI-generated feedback. Beyond this, teachers also holistically evaluated the feedback, taking into account the extent of their refinements in structured review table. Additionally, the study also explores teachers' perspectives on the advantages, limitations, and future use of AI feedback, along with their concrete suggestions for its improvement.

Thus, this study seeks to answer the following three research questions, which aim to contribute to the effective and responsible implementation of AI-generated feedback with teacher involvement.

RQ1. How do teachers refine AI-generated feedback with their own expertise and insights?

RQ2. How do teachers evaluate the accuracy of AI-generated feedback for both errors and content in student essays?

RQ3. What are the advantages and challenges teachers experience in refining AI-generated feedback and their perspectives on its future classroom use?

By answering these research questions, the present study intends to make valuable contributions to the literature in the following three ways:

First, the present study examines how teachers refine AI feedback, and explore how teachers evaluate the accuracy of AI-generated feedback after these refinements. Thus, it also explores the teacher's role as a filter and the extent of teacher filter for error and content feedback. Through this analysis, the research aims to provide data-driven insights into teacher practices and decisions, and offers evidence-based recommendations for the effective and ethical integration of AI tools in classroom settings.

Second, the study contributes to the development of pedagogically valid prompt design based on teachers' correction patterns and rationales. By analyzing how teachers modify AI feedback and understanding their reasoning for these changes, the study provides guidelines for creating more effective prompts that generate outputs aligned with educational goals. This addresses a critical gap identified by Hirunyasiri et al. (2023), who emphasized that effective prompt design is essential for generating accurate and task-appropriate AI outputs.

Third, the research informs future AI design and policy by documenting both the strengths and shortcomings of AI feedback. By identifying where AI excels, where it falls short, and how teachers prioritize feedback, it establishes a baseline for AI developers to design outputs that better align with pedagogical goals. It also supports teachers' AI literacy by making these strengths and weaknesses more visible.

Replicating this study with each new AI release could gradually improve teacher-like accuracy in AI-generated feedback and generate insights that better align AI outputs with student needs and instructors' pedagogical priorities. Over time, these advancements could make AI technology more responsive to both teachers and students.

1.4 Limitations

The limitations of this multiple holistic case study are numerous, although some choices were intentionally made to keep the focus narrower for in-depth qualitative analysis. To start with, the AI models used in this study were pro versions of Gen-AI tools (GPT-4o, Gemini-Advanced, and Claude 3.5 Sonnet). Therefore, the outcomes are only applicable to the specific models available at the time of this research.

Although this study employed pro-plan versions, it is important to note that the models were not specifically designed for educational purposes and hence providing pedagogically sound feedback. Their responses to user prompts were also shaped by other technical constraints, which affected their ability to handle multi-step or highly detailed instructions. While AI models claim to retain large amounts of information in memory and token limits (input + output) appear sufficient for this study's needs, it was observed that the accuracy and quality of responses declined during longer chat sessions and multi-step tasks. This issue became evident during training and while generating feedback for nine papers within the same session.

This outcome pointed to the gap between what the models were theoretically capable of and how they actually performed when handling complex tasks. Therefore, limitations related to memory, instruction complexity, or task length should be recognized as significant barriers to the quality and quantity of AI-generated feedback. This becomes particularly important when detailed and comprehensive feedback is required, as in this study.

Another limitation is the small number of participants ($n = 8$), which may not fully represent teachers' filtering practices and conceptions of AI-generated feedback. However, this small sample size allowed for more manageable interpretation and helped prevent repetitive findings. At the same time, due to the cognitive load involved in refining AI-generated feedback, teacher oversights and varying subjective interpretations are likely to occur during teacher refinements, despite multiple measures such as calibration and standardization. However, when viewed positively, such subjectivity in interpreting and refining AI-generated feedback can enrich the findings of this qualitative study by adding depth to the interpretation of it. It also allows for the exploration of teachers' views within the constraints of this sample size, helping to identify commonalities.

The context in which the study was implemented presents another limitation for generalizability. The use of metalinguistic feedback, implemented as part of institutional practice, may be seen as a threat to the generalizability of AI's feedback capabilities. However, it is important to note that metalinguistic feedback is widely used in other preparatory English classrooms both in Turkey and internationally. Therefore, the findings can still contribute meaningfully to the field.

Besides the feedback mode, the number of training samples could be seen as another setback, as it may affect the training outcomes. However, the collection of essays employed for training were selected from B1-level student papers archived by the university's testing office. In selecting the samples, the testing office considered my research needs and provided the best possible representative subset, including various types of errors as well as content-related feedback issues that had been previously provided feedback by teachers.

Finally, the study's scope was deliberately limited to papers of the same proficiency level (B1) and genre (cause-and-effect essays), with a total of ten papers, including one pilot. While this may limit generalizability, this focused scope enabled a more targeted analysis of AI's performance.

1.5 Definition of Key Terms

This section presents the definitions of essential terms that will be referenced throughout the study. The definitions are outlined below:

Generative Artificial Intelligence (Gen-AI): Generative Artificial Intelligence (Gen-AI) refers to computational systems that employ deep learning architectures to create novel content across multiple modalities, including text, images, audio, and video, by learning patterns from extensive training datasets (Google, 2024). A well-known Pioneer example of this is ChatGPT, which can produce human-like outputs based on user-prompted interactions and provided context. Other Subsequent examples include Google's Gemini and Claude.

AI Chatbots: AI Chatbot is a conversational technology designed to mimic human interactions using text or voice. It generates human-like language and responds to user queries and prompts. While often considered a new technology, its origins can be traced back to the 1960s, with early examples like ELIZA, developed by Joseph Weizenbaum,

which simulated a therapist by reflecting user input, and ALICE, developed by Richard Wallace, which used pattern-based rules for more advanced conversations. A more recent example is Apple's Siri, which incorporates some Chatbot features. These chatbots, however, are incapable of understanding complex tasks. More technologically advanced chatbots have since emerged, capable of handling complex and diverse tasks while engaging in human-like conversations. Notable examples include GPT, Gemini, and Claude (Labadze et al., 2023).

Large language models (LLMs): They can be defined as advanced deep learning models trained on massive text datasets to comprehend and generate language resembling that of humans. This human-like capability arises from their ability to learn from the provided context and produce relevant responses to user queries (Devlin et al., 2018; Yu et al., 2023).

AI-generated feedback: In the context of this study, AI-generated feedback refers to feedback produced by AI in response to the prompt developed after iterative trials by the researcher. After trying out three different platforms (GPT-4o, Gemini-Advanced and Claude 3.5. Sonnet, this feedback was obtained which involves three primary outputs for student papers to emulate teacher-like feedback: (1) an error-incorporated paper, (2) content feedback, and (3) a detailed error table with error descriptions written in a constructive and guiding tone. The aim is to create teacher-like metalinguistic feedback that encourages students' self-discovery of their errors and perform self-correction.

Teachers: In this research context, "teachers" refers to preparatory class instructors working in classrooms within the institution. Their role in this research involves refining AI-generated feedback, both error-coded and content-related, on student essays using a metalinguistic feedback approach, as proposed by Truscott (2007). Following this approach, their review involves checking and refining whether the AI uses appropriate error codes accurately, provides error descriptions in an appropriate tone, avoids offering explicit solutions in content feedback, and delivers content feedback that is accurate, concise, and targeted. The feedback is also expected to consider both strengths and areas needing improvement, all phrased in a constructive and guiding tone. Additionally, during the qualitative analysis, teachers are anonymously referred to as "T" followed by a number (e.g., T1 to T8) to ensure the confidentiality of research participants.

Color-coded review form: For the purposes of the study, teachers performed color-coded reviews on the student papers shared in Word documents accessed via OneDrive. They applied a predefined color-coding system to highlight incorrect, redundant, or overlooked AI-generated feedback on sample student essays. Thus, this method enabled traceability of teacher modifications for error-coded feedback.

Structured review table form: In addition to the color-coded refinements, teachers completed structured review tables shared with them in a spreadsheet format and accessed via Google Drive. These tables were used by teachers to record their acceptance levels for each feedback item, select rationale codes from a dropdown list, and provide notes explaining their changes.

Accuracy Evaluations: In this study, teachers employed an accuracy evaluation rubric to assess the accuracy of AI-generated feedback on student writing. The rubric provided a structured framework for evaluating the accuracy of feedback across several dimensions holistically, not only focusing on whether the AI correctly identified the error, but also on its pedagogical accuracy, including whether the explanation was clear, the tone was guiding and constructive, and whether the feedback offered actionable guidance to support student improvement.

Token Limit (AI Token Limit): This refers to maximum number of tokens an AI model can handle in a single prompt and response process. It includes both the user's input and the AI's output. For example, Claude 3.5 Sonnet supports up to 200,000 tokens, while GPT-4o can process up to 128,000. If this limit is exceeded, the AI may cut off the output or lose part of the context.

Readability Index: In this study, it was important to evaluate content feedback for readability appropriate to B1-level language proficiency. Therefore, a readability index was calculated; a numerical score used to estimate how easy or difficult a text is to read. Two readability indices were used in the present study. In Text Inspector web tool (Weblingua Ltd., n.d.), readability measures such as Flesch-Kincaid, Gunning Fog, and SMOG are calculated based on factors like sentence length, word length, and syllable count. Among these, the results of the Flesch-Kincaid index were employed to assess the complexity of the feedback in terms of B1-level readability.

In addition to Text Inspector, the Oxford Text Checker (Oxford University Press, n.d.) was also employed to evaluate whether the vocabulary used in the content feedback was

accessible to students. Unlike Text Inspector, the Oxford Text Checker does not calculate readability scores, but it supports readability analysis by identifying the CEFR level of each word (A1 to C2). Thus, it helped determine whether the vocabulary used in the content feedback remained within the boundaries of the target proficiency level of the learners.



2 REVIEW OF LITERATURE

The literature section begins by providing a historical and pedagogical overview of how different feedback sources have evolved over time, from traditional teacher-led feedback to AI feedback. By delving into these theoretical frameworks, the review provides a clearer picture of how feedback practices have shifted and what factors have contributed to the growing interest in AI-supported feedback. It explores both the theoretical arguments in favor of integrating AI into writing instruction for feedback purposes and the concerns raised about its pedagogical appropriateness and potential drawbacks. In doing so, this section presents a balanced view of the theoretical landscape, showing how both supportive and critical perspectives have shaped current discussions on the role of AI in writing feedback.

2.1 A Historical Look at Feedback in L2 Writing

As the approaches to learning evolve from teacher-centered to student-centered with the involvement of technology, the practices of feedback have also undergone a similar transition over time, shifting from teacher-centered approaches to machine-based feedback, and most recently, to generative AI feedback providing autonomous learning opportunities to students. In this shift, each and every approach in the making of learning history has uniquely contributed to technology-assisted feedback practices, culminating in the development of generative AI feedback. The table 2.1. below presents a chronological timeline of these learning approaches and provides a summary of their impact on the feedback practices.

Table 2.1

Key Theoretical Approaches and Their Contributions to Feedback Practices

Theoretical Approaches	Key Scholars	Core Contributions to Feedback Practices
Behaviorist Approach	Skinner (1954)	Focused on reinforcement using positive or negative stimuli to change behavior. Teachers provided corrective feedback centered on grammar, structure, and syntax, symbolized by marking errors with red ink.
Constructivist Approach	JPiaget (1971, 1985)	Emphasized knowledge construction through guided discovery. Feedback shifted to prioritizing content over grammar, encouraging creativity and personalized responses.

Table 2.1 (continued)*Key Theoretical Approaches and Their Contributions to Feedback Practices*

Theoretical Approaches	Key Scholars	Core Contributions to Feedback Practices
Cognitive Process Theory	Flower & Hayes (1981)	Viewed writing as a recursive process involving planning, translating, and reviewing. Feedback became process-oriented, focusing on content development and logical flow before addressing grammar in later drafts.
Social Constructivist Approach	Vygotsky (1978)	Highlighted the importance of social context and interaction in learning. Feedback expanded to include peers as “More Knowledgeable Others” (MKO), enabling collaborative learning and reducing teachers' workload.
Generative AI Feedback	Banihashem et al., 2024; Farrokhnia et al., 2023	Transitioned from grammar-focused machine feedback to communicative Generative pre-trained AI tools like ChatGPT and Claude, which can offer personalized and context-sensitive feedback.

As summarized in Table 2.1 above, the historical progression of feedback approaches has been shaped by the dominant learning theories of their respective periods. Their influence can also be seen in how feedback is provided and communicated. These approaches, along with the evolution of feedback, are explained in detail and delineated according to their respective time periods in the sections that follow.

2.1.1 Behaviorist approach to Feedback (1950s)

Back in the 1950s, the behaviorist approach was the dominant learning theory. The foundations of behaviorism were laid earlier by Watson (1913), but it was Skinner’s (1954) Operant Conditioning theory that shaped classroom applications of feedback and learning behavior. His theory situates learning as a process in which behavior is strengthened or weakened based on its consequences, particularly through the use of positive or negative reinforcement and punishment. In this respect, learning is strictly seen as habit formation, and errors are less tolerated and not approached constructively.

Feedback practices were also shaped by this perspective, and Skinner’s Teaching Machine became an emblem of behaviourist learning and feedback. The device, known as the teaching machine, demonstrated how behaviourist learning and feedback occur. This machine reinforces learning through immediate correction of wrong answers and provides repetitions to help learners correct those errors. Incorrect answers must be

corrected before moving on. In this regard, feedback simply indicates that the answer is wrong, without explaining why it is wrong or what the correct answer is.

The teaching philosophy behind this machine was to break incorrect habits through behavior shaping, as learners gradually adjust their behavior based on the feedback they receive before these errors become ingrained or fossilized in students' learning. As an error-focused model, this approach emphasizes identifying and correcting mistakes, but it neglects the constructive or guiding aspect of feedback. As a result, it can be said that this approach often lacks motivational value for learners and does not aim to encourage students to engage in self-correction.

2.1.2 Constructivist approach to feedback (1970s-1980s)

As a reaction to the behaviorist approach's lack of emphasis on motivation and constructive guidance the 1970s and 1980s witnessed a shift toward more constructivist feedback practices that emphasized learning through experience. That is to say, students' mistakes began to be viewed as opportunities for learning rather than as signs of failure. One of the most influential scholars associated with this perspective was Jean Piaget. His theory focuses on cognitive development stages that influence learning, which are: the sensorimotor, preoperational, concrete operational, and formal operational stages (Piaget, 1971; Piaget, 1985).

As can be seen from the description of the theory, it is not directly related to writing. However, just like the behaviorist theory, it has influenced writing instruction and feedback provision through its pedagogical philosophy. Piaget's theory, which illustrates the staged growth of an individual's thinking based on age, suggests that learning is not a one-time experience. Rather, it involves a gradual process consisting of various developmental stages and improves through experience. In other words, learning occurs in different phases of life, and individuals discover rules or construct understanding by actively engaging with their environment, an idea also supported by the notion of guided discovery (Brooks & Brooks, 1999; Lutz & Huitt, 2004). This constructivist approach, therefore, defines the teacher's role as that of a facilitator who avoids correcting student errors directly. Instead, teachers guide students to discover and correct their own errors and learn through trial and error.

Influenced by the constructivist approach, Graves (1983) explained how to guide students through feedback on writing by introducing errors in stages, in order to facilitate the

learning process. In his view, content should be prioritized over grammatical correctness, especially in the early stages of writing, as excessive focus on grammar can demotivate learners and hinder them from developing their unique voice as writers. He also advocated for the use of personalized feedback rather than doing merely mechanical correction, emphasizing that each student is unique and has different learning needs to prioritize.

2.1.3 Cognitive process theory and feedback (1980s)

Later, the cognitive process theory emerged, influenced by Piaget's ideas. This approach changed the traditional view of writing as a linear activity and instead emphasized writing as a process, which also influenced feedback practices by encouraging support throughout the different stages of writing. In their seminal work, Flower and Hayes (1981) defined writing as a cognitive activity that includes recursive stages such as planning, translating, and reviewing. This also shifted the focus away from a product-based approach, which often led students to view feedback as a grade-oriented practice. Instead, feedback became more concerned with improving students' writing across different stages and encouraging comparisons during and after the revision process (Smith, 2015). Hattie and Timperley (2007) coined the term "incremental feedback" to describe this process-oriented approach. From their perspective, feedback in the early stages of writing should focus on content development, particularly the organization of ideas and the logical flow of the text. In later stages, feedback should shift toward language use and style, with an emphasis on grammatical accuracy in the final draft.

2.1.4 Social constructivist approach to feedback (1978, onwards)

Around the same time, the social constructivist approach emerged in response to the limitations of behaviorist and early cognitive theories, both of which viewed learning as an individual endeavor. Instead of seeing learning as an isolated process, Vygotsky's social constructivist theory (1978) emphasized the importance of social context, positioning learners as active participants in the construction of knowledge. Unlike previous learning theories, Vygotsky's work places a strong emphasis on social interaction and the role of socialization in the learning process.

This approach to learning has also influenced how feedback was delivered to students, especially as the role of the teacher has changed to that of a facilitator, and feedback was no longer viewed as something only to be provided by a teacher. Vygotsky's theory did not underestimate the role of teachers, but it emphasized that assistance can come from

other external sources within the Zone of Proximal Development (ZPD). The ZPD refers to the gap between what a learner can do independently and what they can do with help from someone more knowledgeable. Vygotsky suggested that students learn better when they receive assistance from a knowledgeable other, which does not have to be a teacher only, but extends more comprehensively to include anyone knowledgeable. This can be parents, teachers, or peers who possess greater knowledge than the learner.

This theory emphasizes the role of interaction in that it claims students can surpass the limitations of their current knowledge and reach the next level without feeling frustrated when they receive help from a more knowledgeable other (MKO). With this idea in mind, many studies began to investigate whether peers can be an alternative to teachers in providing feedback, with the hope of reducing the feedback burden on teachers (Er, Dimitriadis & Gašević, 2021; Kerman et al., 2024; Lu, Zhu & Cheong, 2021; Noroozi et al., 2023).

During this time, technology was not as advanced as it is today, and peers were evaluated as an alternative to teachers, since the learning process was experienced in class and students were considered part of this social community of learning.

2.1.5 Transition to machine feedback and generative AI chatbots (2010s-Present)

With the changing educational landscape fueled by advancements in technology, the teacher-dominant mindset in feedback practices has given way to technology-mediated feedback. The shift to technology-mediated feedback was caused by several innovations in the field of feedback provision.

The first innovation marking the emergence of machine feedback, was supported by AI (hereafter referred to as AWCF tools), such as Grammarly and Wordtune. These AWCF tools primarily use natural language processing (NLP), as well as statistical and prediction-based algorithms to produce feedback (Wilson & Roscoe, 2020). In other words, NLP allows computers to understand and work with human language based on statistics and predictions, since there was no two-way interaction with the user to confirm and verify the tool's findings.

To describe earlier machine feedback capabilities further, one of the most well-known and still widely used Automated Writing Correction and Feedback (AWCF) tools today, *Grammarly*, has imparted key information on its website regarding its capabilities in

providing feedback. This information covers what Grammarly is capable of, as well as what it is not. According to that, Grammarly is limited to grammar correction and does not provide feedback on contextual issues (Grammarly, 2024).

This suggests that AWCF feedback AI-supported tools initially focused heavily on grammar and cannot produce context-sensitive feedback owing to the constraints of their algorithms. However, these tools have evolved, shifting from basic grammar-focused proofreading tools to generative AI models known as Large Language Models (LLMs), commonly referred to as Chatbots. These models differ significantly from their predecessors, as they are designed to generate conversational, human-like feedback (Yan et al., 2024). In contrast to earlier AI tools, these models claim to understand context and tailor feedback according to student needs because they can be trained with large-scale data and human user input.

These Chatbots are referred to as generative tools (Banihashem et al., 2024; Farrokhnia et al., 2023), and today, various Gen-AI platforms such as ChatGPT, Gemini, Claude, and many others have emerged as potential alternatives for providing feedback to support and complement teacher feedback.

2.2 Conceptual & Theoretical Frameworks for AI feedback

The theoretical foundations of AI-generated feedback did not emerge overnight; they are shaped by decades of educational research and built upon a long history of insights into how people learn most effectively. Therefore, AI-generated feedback should not be perceived as disconnected from well-established learning theories such as constructivism, sociocultural theory, cognitive learning theory, formative assessment theory, and other influential learning hypotheses.

These foundational theories have long informed the design and implementation of educational technologies. Hence, it would be wrong to say the establishment of AI also rests on these established theoretical perspectives. Table 2.2 summarizes the key theoretical frameworks that support the pedagogical relevance and applicability of AI-generated feedback. Furthermore, it presents recent findings from studies that reinforce this theoretical alignment and demonstrate AI's potential to function effectively as a feedback provider in educational contexts.

Table 2.2*Theoretical Frameworks Supporting AI Feedback and Their Alignment with Present Literature Findings*

Theoretical Framework	Past Scholars	Current Literature Findings Aligned with Past Theories (Supportive of AI Feedback)
Behaviorist Theory	Skinner (1953)	AI aligns with the notion of reinforcement because it allows for positive feedback (e.g., praise) or negative feedback (e.g., corrections). AI uses these principles to improve output quality and encourage desired behaviors (Dai et al., 2023).
Social Constructivist Theory	Vygotsky (1978)	AI acts as a more knowledgeable other (MKO) and can provide support within the zone of proximal development (ZPD). AI adapts to the learner's level and provides personalized feedback, aiding both students and teachers (Urooj & Li, 2021; Evmenova et al., 2024).
Comprehensible Input Hypothesis	Krashen (1985)	AI grades its language to be comprehensible and adjusts it slightly beyond the learner's current proficiency level ($i+1$), acting as a technological MKO (John-Steiner & Mahn, 1996).
Interaction Hypothesis	Long (1996)	AI facilitates meaningful interactions through conversational interfaces, enabling negotiation for meaning and clarifications, akin to human interaction.
Output Hypothesis	Swain (1985)	AI delivers immediate and detailed feedback to help learners notice and correct gaps in their language production, enhancing autonomy and engagement (Li, 2020; Wang, 2019; Zhang & Hyland, 2018; Koltovskaia, 2020).
Formative Assessment Theory	Black and William (1998) Boud (2000)	AI supports formative assessment by providing consistent, detailed, and constructive feedback during the writing process. This approach supplements teacher feedback by addressing errors and offering personalized suggestions (Vashishth et al., 2024).

In the following sub-sections, the theoretical alignment of Generative AI feedback with past frameworks is explained in detail.

2.2.1 Behaviorism and AI: is AI behaviorist?

Although AI is a product of technological advancements, it has been developed and trained based on well-established theories from the past. For instance, Skinner's (1953) Behaviorist Theory of Learning deeply affected AI algorithms because AI acts based on data-driven, and AI learning also requires behaviorist principles. That is, AI learns and improves its output through exposure to large amounts of input data, corrections, and user interactions, reflecting a process of reinforcement-based learning. Three main principles are used in machine learning: supervised learning, which involves labeled data to help the system recognize patterns; unsupervised learning, which uses unlabeled data to identify patterns and form groupings; and reinforcement learning, which involves trial and error

and continuous refinement of its responses through rewards and penalties (Holmes, Bialik, & Fadel, 2019).

In the provision of feedback to students, it was also influenced by behaviorist approach to feedback, so it can provide “immediate” correction with positive or negative reinforcements. In this sense, AI can act much like Skinner’s Teaching Machine, which can be seen as an early precursor or conceptual ancestor of today’s AI-powered feedback systems (Holmes, Bialik, & Fadel, 2019). These systems deliver immediate feedback, reinforce correct answers with positive feedback, and respond to incorrect answers by asking retry. In the recent study of Dai et al. (2023) acknowledged this role, stating that AI Chatbots can follow the principles of reinforcement principle of behaviorist theory while providing writing feedback. The findings revealed that AI could not only provide corrective feedback, but also include praise for correct responses. In line with the behaviorist approach, this finding indicated that AI can be behaviorist in its feedback style, and reflect this reinforcement-based structure. However, AI is not purely behaviorist because they can include It includes some explanations (i.e., metalinguistic guidance), and reflects cognitive and communicative aspects of learning too.

2.2.2 Vygotsky’s social constructivist theory: is AI the new MKO?

Later generations of AI, particularly Chatbots, were influenced a lot by Vygotsky’s social constructivist theory (1978). This theory emphasizes the importance of social interaction with others, such as peers, teachers, and members of society, in the process of learning. Chatbots were designed to embody this kind of interaction with users. As a result, core concepts from Vygotsky’s theory, including the Zone of Proximal Development (ZPD) and the More Knowledgeable Other (MKO), have significantly shaped the design of these interactive AI systems.

As a matter of fact, when Vygotsky first introduced the concept of the More Knowledgeable Other (MKO), it referred to any person who was more knowledgeable about a subject than the learner, such as a teacher, parent, or peer. Hence, at the time, it was not ascribed to a technological tool. However, in later years, John-Steiner and Mahn (1996) elaborated on and expanded Vygotsky’s conceptualization of the MKO, taking into account the growing role of technology in education. As a result, technological tools began to be considered as potential embodiments of the MKO in the context of a modern, technology-driven world.

Similarly, research has also highlighted the connection between AI and the Zone of Proximal Development (ZPD). According to study conducted by Urooj and Li (2021), AI, being data-driven, can assist students in learning without overwhelming or under-challenging them by keeping the support within the ZPD and properly adjusting to the student's level. Aside from giving support to students within the level of ZPD, AI can also aid teacher to deliver feedback for struggling writers by detecting their areas of struggle (Evmenova et al., 2024).

2.2.3 Krashen's comprehensible input: is AI feedback comprehensible?

Krashen (1982, 1985) proposed the Input Hypothesis as part of his theory of second language acquisition, along with the Monitor Hypothesis. His concept of the Comprehensible Input Hypothesis is especially useful for understanding how interaction with AI should be structured and how AI output should be delivered. According to this hypothesis, input is most effective for learning when it is slightly above the learner's current level of proficiency, what Krashen refers to as "i+1." This is defined as comprehensible input: input that is neither too difficult nor too simple for the learner.

Comprehensibility of input must be ensured in what AI delivers to students, and in a study of Jin et al. (2025) it was highlighted that AI can offer feedback across varying levels of English proficiency based on the Common European Framework of Reference for Languages (CEFR). However, this rests on your prompt. That is, to generate comprehensible input slightly above their level without overwhelming them, it is critical how AI is prompted and trained to generate appropriate output. This means that the input provided by the user/ trainer should be clear and understandable because this can influence the generation of output by AI. Users need to formulate prompts using simplified and direct language to ensure that the generated feedback remains accessible and useful for students. When the input provided in an AI prompt is too complex or vague, the resulting feedback often fails to align with CEFR levels. This issue was also highlighted in the study by Jin et al. (2025), who found that although AI is capable of producing CEFR-aligned, comprehensible input, simply using prompts such as "write at A2 level" is not sufficient. To address this, they employed the CEFR Readability Classifier (CRC) tool to evaluate and guide the AI's output. They also provided example sentences that demonstrated the desired proficiency level to help generate appropriate,

level-aligned feedback. This suggests that achieving the desired output requires more than a single prompt, it involves fine-tuning with external tools and well-structured input.

2.2.4 Interaction hypothesis and AI as a conversational partner

A more precise answer to the question of whether AI can communicate in a level-appropriate way for the respondent's level can be found in Long's Interaction Hypothesis (1996). According to this theory, interaction with an interlocutor facilitates language acquisition. In a dialogue, an individual adjusts their language or uses interactional strategies, such as clarification requests or confirmation checks, to communicate their message more clearly. Supported by this theory, AI chatbots, with their interactionist interfaces, are considered capable of facilitating such dialogues and negotiations of meaning with users. In cases of communication breakdowns with AI, it is possible to clarify misunderstandings and pursue further queries, as if communicating with a person. Additionally, it can be observed that AI begins to adapt its language as the input increases in the same chat and refers back to earlier interactions in its responses.

2.2.5 Output hypothesis and contribution of AI to noticing and self-correction

Swain's Output hypothesis (1985) can also be associated with the feedback output provided by AI and ways to improve it. This theory argues that, aside from input, output is an essential component of language acquisition because learners can notice their gaps only when they produce language and try to fill in those gaps and breakdowns when they realize their messages are not being conveyed. This relates to AI because users can check the output produced by AI and provide more input or context to receive a more desirable response. Meanwhile, when discussed in relation to student learning, Li and Xu (2020) indicated that AI can deliver feedback in a short time, helping students notice their gaps immediately and correct their grammatical mistakes without delay. This also suggests that AI can provide feedback instantly and tailor it to individual needs, enabling learners to correct and adjust their language more easily. In agreement with this view, several other studies have highlighted that AI tools can enhance student autonomy and engagement, leading to improved writing outcomes (Koltovskaia, 2020; Zhang & Hyland, 2018).

2.2.6 AI as a tool for formative feedback

Formative assessment plays a vital role in helping students improve their writing, and AI is increasingly being explored as a tool that can support this process. Black and Wiliam (1998) argued that formative assessment is essential for teachers to fully understand student difficulties and address them, as well as to foster student progression through consistent feedback on errors. Boud (2000) also noted that while teachers often view providing feedback as formative, assuming it will automatically lead to student learning, in real-life contexts, this does not necessarily happen unless a student acts upon and improves their draft based on it. These improvements with the final draft, in return, are also communicated back to the teacher who provides the feedback. This process of “closing the loop” ensures that students truly understand the feedback and genuinely learn from it. And it is only then that teachers can observe how effective their feedback has been.

However, in traditional settings with large student numbers, it is challenging for teachers to manage such individualized feedback loops, as they require in-depth knowledge of each student's profile, errors, and needs (Double et al., 2020). AI, however, shows promise by making personalized and instantaneous feedback loops feasible. By supporting student progression in this way, AI can effectively facilitate the principles of formative assessment encapsulated in process writing, which involves providing ongoing, supportive, and constructive feedback throughout the writing stages.

AI can particularly assist teachers in providing individualized feedback to students based on the analysis of student work, allowing for formative assessment (Vashishth et al., 2024). This individual feedback can be useful and timesaving for teachers, as it provides data-driven insights regarding students’ struggling points and helps address writing problems. In turn, this allows teachers to adapt their feedback strategies more effectively.

2.3 Overview of Theories that Non-supporting AI Feedback

The previous section focused on theories that can support the use of AI for feedback provision. In contrast, this section provides dwells on opposing views towards use of AI as a source of feedback. It mainly focuses on Noam Chomsky's hypotheses and Searle’s Chinese Room Argument for their implications. Table 2.3 presents a summary of theoretical perspectives that argue against the use of AI for feedback with some implications.

Table 2.3*Theoretical Frameworks Supporting AI Feedback and Their Alignment with Present Literature Findings*

Theoretical Framework	Key Scholar(s)	Core Argument	Implications for AI Feedback
Innateness Hypothesis (1965)	Chomsky	Language acquisition is innate; true intelligence involves creativity and contextual reasoning that AI lacks.	AI cannot replicate innate human language understanding; feedback may lack depth or learner relevance.
Poverty of Stimulus (1980) & Universal Grammar (1986)	Chomsky	Language is learned with limited input, suggesting an internal system not replicable by AI.	AI's feedback might miss intuitive language cues or over-rely on surface-level patterns.
Chomsky's Critique of AI's Predictive Language Model (Chomsky, 2023)	Chomsky	AI is based on statistical prediction, not true understanding. It fails to interpret complex syntax and lacks reasoning and awareness of meaning.	AI feedback may misinterpret ambiguous sentences, leading to misleading or shallow suggestions.
Chinese Room Argument (1980)	Searle	AI manipulates symbols without real comprehension or awareness of semantics.	AI-generated feedback may appear intelligent but lacks true insight, raising concerns about pedagogical reliability.

2.3.1 Noam Chomsky's innateness hypothesis

Chomsky, the father of modern linguistics is a noteworthy reference point for understanding the limitations of AI in comparison to human intelligence. In fact, his insights, which date back to earlier years, accounted for first language acquisition, but he also employed his theory to explain why AI might not be a highly trustworthy source for providing feedback on student writing papers.

Very recently, Chomsky (2023) has published a manifesto expressing his disbelief in AI's potential to replicate human-like language capabilities in *New York Times Magazine*. In this manifesto, he presented his strong theoretical framework, the *Innateness Hypothesis*, to argue against the use of AI in language contexts. Chomsky (1965) asserted in this hypothesis that the human mind cannot be replicated by AI, a view central to his Innateness Hypothesis. This hypothesis suggested that a child learns language instinctively and naturally, building knowledge of grammar with very limited exposure, a phenomenon he named the *Poverty of Stimulus* (Chomsky, 1980). Based on this premise, he also proposed the existence of *Universal Grammar*, which posits that

language capacity and the ability to learn grammar are innate systems wired into humans (Chomsky, 1986). This is mainly where Chomsky and AI developers diverge.

Besides, in *New York Times Magazine*, Chomsky (2023) explains his view further by juxtaposing human thinking with AI thinking. He argued that AI lacks true intelligence, as it relies on a data-driven learning system based on prediction. To be more precise, Chomsky's understanding of true intelligence entails more than just predicting or describing things based on data. True intelligence in his term essentially requires cognitive flexibility and creativity to explain why things are the way they are and to generate different reasonings or justifications for an issue, as humans can (Chomsky, 2023, paras. 2–3).

Therefore, he contended that AI cannot explain the principles of English syntax and cannot compete with the complexities of human thought, and he provided the following example to demonstrate the lack of contextual awareness. For instance, Chomsky provided the sentence “John is too stubborn to talk to” as an example to illustrate that AI lacks the contextual understanding required to interpret such structures accurately. The accurate interpretation of this line would be, “John is too stubborn to be reasoned with,” but AI misinterprets the meaning conveyed in the sentence and predicts the meaning as, “he refuses to talk” (Chomsky, 2023, para. 7). This misinterpretation, Chomsky claims, highlights AI's inability to grasp more complex grammatical structures and the deeper layers of meaning present in natural human language.

Chomsky's explanations mentioned above highlight the shortcomings of AI in terms of true intelligence, that is understanding of human language and reasoning. When these ideas are applied to feedback practice, it would not be wrong to infer that AI-generated feedback cannot match the quality and precision of human thinking.

2.3.2 Searle's Chinese Room Argument

The Chinese Room Argument a thought experiment proposed by the American philosopher Searle in his 1980 journal article, *Minds, Brains, and Programs*, and the argument can be applied to AI because it aims to show that computers are not capable language. In the experiment, Searle, posits a scenario where a man who does not know Chinese is locked in a room and given a rulebook that allows him to manipulate Chinese symbols and produce appropriate responses, despite having no understanding of the Chinese language. Although it appears to the observer in the other room that this person

understands Chinese, this person in fact has no genuine understanding of context and semantic meaning. This experiment demonstrates that although AI may appear to emulate human thought or intelligence and produce human-like language, it does not actually understand language in the way humans do. AI operates on pre-trained data of knowledge of syntax, and however, unlike humans, it does not process language intuitively based on the demands of context. This idea also resonates well with Chomsky's skeptical perspective on AI's intelligence and language understanding.

2.4 Defining the Characteristics of Effective Feedback in L2 Writing

Informed by the earlier theoretical discussion, this section examines the conceptual framework used to define the key characteristics of effective feedback in L2 writing. It draws on a range of studies to identify what makes feedback meaningful, actionable, and pedagogically sound. Following that, a cross-comparative analysis of various feedback sources, including teacher, peer, automated machine, and AI feedback in terms of their ability to meet these characteristics. Toward the end of the section, a summary table is provided to compare AI feedback to both teacher and peer feedback. This comparison highlights the potential advantages and existing shortcomings of AI feedback, offering a solid foundation for the present study's exploration of its pedagogical suitability for feedback delivery.

2.4.1 Broad definition and key features of effective feedback

An extensive body of research has explored the essential characteristics of effective feedback and offered various definitions to describe it. As a result, the literature does not present a single, universally accepted definition, since feedback effectiveness encompasses multiple coexisting components. Consequently, scholars have proposed different, yet often overlapping, lists of key characteristics of effective feedback, along with guidelines for teachers on how to deliver it effectively. The supporting literature on the traits of effective feedback, along with various descriptions provided in Table 2.4.

Table 2.4*Supporting Literature on the Key Characteristics of Effective Feedback*

Characteristic	Authors and Year	Description
Three Aims of Feedback	Sadler (1989); Hattie & Timperley (2007); Black & Wiliam (2009)	Good feedback should address three aspects for student learning: what is expected, the gap between their work compared to a standard one, and what the student needs to do to improve the work.
Understandable Feedback	Nichol (2010)	Feedback should be expressed in a language that students can easily understand.
Contextualized & Balanced Feedback	Nichol (2010)	Feedback should be tailored to fit the assessment criteria, framed to meet learning outcomes, mention both positive aspects, and highlight areas needing improvement.
Personal Feedback	Nichol (2010); Brookhart (2011)	Feedback should acknowledge the student's previous performance by referring to their earlier work and commenting on improvements or areas to improve.
Praise the Work, Not the Student in Feedback	Brookhart (2011)	Comments should praise the student's work specifically, rather than giving general praise to the student.
Refer to Previous Work & Provide Few Points in Feedback	Brookhart (2011)	Feedback should compare the current work to previous work to personalize the feedback and prioritize a few key areas for improvement to avoid overwhelming the student.
Simple Vocabulary & Language in Feedback	Brookhart (2011)	Feedback should use simple language and short sentences to ensure that the student understands it.
Check Student Understanding for Feedback	Brookhart (2011)	Ask students to identify the most important problem or missing element in their work to check their understanding of the feedback.

As can be seen, scholars have mentioned various components for characteristics of effective feedback. To start with fundamentals of effective written feedback, Sadler (1989) emphasized three key components that need to coexist for students to understand, interpret, and take action. These conditions are:

1. Understanding the expected standard.
2. Identifying the gap between the quality of their work and the expected standard.
3. Knowing the steps needed to close the gap.

Similar to Sadler's (1989) ideas, Hattie and Timperley (2007) and Black and Wiliam (2009) proposed evidence-based feedback models supported by meta-analysis results. Their model defines three essential qualities of effective feedback through three guiding questions:

1. Where are you going? (your goal as a writer)
2. How are you going? (The quality of your work; things you did good or bad)
3. Where to next? (Areas needing improvement)

These three questions are also referred as the three aims of feedback:

1. Feed up (clarifying goals).
2. Feedback (assessing current performance).
3. Feed forward (guiding improvement).

When these elements are missed from feedback, feedback becomes overtly vague or generic, with comments like 'good work' or 'very clever ideas', providing insufficient guidance for students to act upon. Moreover, such feedback may lead students to focus on themselves rather than on improving their work (Hattie & Timperley, 2007).

In addition to these three building blocks of feedback, scholars also identified several ground rules in order to tailor the feedback responsive to student needs. Nichol (2010) identified the following seven guidelines:

1. *Effective feedback is understandable.* It should be expressed in a language that students will understand.
2. *Effective feedback should be selective.* It should include commenting in reasonable way, focusing on two or three things that the student can action.
3. *Effective feedback should be specific.* It should clearly explain what the point is for students to take action to correct his/her mistake.
4. *Effective feedback should be timely.* It should be provided timely to improve students' next work. That is, the feedback should be provided before student forget his/her work.
5. *Effective feedback should be contextualized.* It should be tailored to fit with assessment criteria and framed to meet learning outcomes expected from the students.

6. *Effective feedback should be balanced.* It should mention both positive things and the areas need improvement. If something negative is mentioned, it should be sandwiched by positive feedback.
7. *Effective feedback should be personal.* It should acknowledge students' previous performance by referring to student's previous work. That is, it should acknowledge and comments on the improvements on the student paper or areas awaiting improvements.

Later, elaborating more on these ground rules, Brookhart (2011) proposed the following aspects for teachers to focus when providing feedback to a student paper:

1. *Focus on one or two strengths and at least one weakness to improve:* Effective feedback focus one or more strengths and at least one thing to improve for the next work. Besides that, she clarifies that effective feedback cannot be comparative to other students' work and hence avoid judgmental tone.
2. *Keep positive tone, clear and specific message:* The effective feedback should not use a commanding language that orders what students should do but should use constructive tone that explain clearly what they can do to improve the work.
3. *Do not praise the student but praise the work:* your comments should not praise the student, but the students' work.
4. *Referring to previous work and establishing link to the present work:* In adjusting feedback, consider referring what students do in the past work by comparing it to the present work. You can describe the strengths or the weakness in students' previous work to make them notice. *Thus, it would be possible to personalize the feedback process and help struggling learners.*
5. *Few points, small steps:* In order to adjust feedback, the other major consideration is to limit the points to a few things, not more. Be selective and choose the thing student needs to focus for the next work. Otherwise, if feedback is extensive, students may not concentrate everything at once. It is important to prioritize some of the things and see if the student progress or not.
6. *Use simple vocabulary:* When feedback is provided, keep the language simple and the sentences short. If the student cannot understand a word in feedback, they cannot do corrections based on the feedback.

7. *Check student understanding*: Ask student what is the most important problem or the missing thing in the work? This can help checking the students' understanding of the feedback.
8. *Commenting sufficiently on the successful student works*: Sometimes, teachers tend to focus on the students struggling while commenting less on the successful student paper. Rather than praising student, praise the work and name the areas of strengths very specifically. And suggest some other things that could be taken in the next work.

2.4.2 Effective written corrective feedback (WCF) in literature

The review of the literature on written corrective feedback (henceforth WCF) has no consensus as to what type of corrective feedback is effective due to varying approaches of corrective feedback. Different scholars have put forward different forms of feedback based on student needs and their instructional goals. For instance, Ellis (2009) identified three widely adopted practices for delivering writing feedback: direct, indirect, and metalinguistic feedback. The distinction between these methods lies in the manner of feedback delivery. Direct feedback provides explicit corrections and solutions to errors, whereas indirect written corrective feedback highlights the error without offering the correct form. Indirect WCF is further divided into four groups based on correction approaches, according to Sheen's (2011) typology. He identified four types of indirect written corrective feedback: indirect WCF (located error), indirect WCF (non-located error), indirect WCF using error codes, and indirect metalinguistic written correction (with a description provided for each type). So, according to Sheen (2011), metalinguistic feedback with codes or descriptions is a subcategory of indirect WCF. It offers cues to student, such as codes or symbol, regarding the nature of the error. Unlike direct feedback, it does not explicitly state the error, but guides students toward self-correction (Li & Roshan, 2019; Luo & Liu, 2017).

The focus of feedback also differs depending on the needs of the students. Ellis et al. (2008) divided them into two types: focused feedback and unfocused feedback. In focused WCF, teacher can selectively address certain errors on the student paper as opposed to trying to fix everything all at once. Thus, students can identify fundamental problems in their writing and do not have to pay attention to everything at once (p. 356). According to them, this approach can be less overwhelming for students.

On the other hand, unfocused WCF involves providing feedback on everything without any selection, which is considered by some as best practice. It is a very common approach among teachers in foreign language writing classrooms (Furneaux, Paran, & Fairfax, 2007). Yu et al. (2020) explained the mismatch between teachers' beliefs and in-class practices by noting that teachers' implementation strategies are shaped not only by their beliefs, but also by contextual constraints such as institutional choices regarding WCF practice, exam policy, and workload.

Additionally, in written corrective feedback (WCF), another parameter to consider when providing feedback is the distinction between local and global error correction. Touchie (1986) explained that local errors are mechanical issues, such as spelling or preposition mistakes, which do not significantly affect the overall understanding of a sentence. In contrast, global errors, if left uncorrected, can obscure the meaning of a sentence. For this reason, Ellis (2009) emphasized that addressing global errors should take priority when providing corrective feedback, as ensuring message clarity is an essential component for the comprehensibility of students' messages.

2.4.3 Teacher, peer, and AI WCF feedback: a comparative review of their characteristics in the literature

The characteristics of teacher, peer, and AI feedback have sparked a heated debate over which feedback mode is the most effective. There are varied interpretations regarding the pros and cons of each feedback mode in literature depending on type the needs of the students and teachers. Consequently, the suitability of each feedback mode is context-dependent, influenced by factors such as learner proficiency, task type, learning objectives, and available resources. In the light of this, the following Table 2.5 summarizes pros (+) and cons (-) of feedback modes: teacher, peer and AI-generated feedback with the existing literature findings.

Table 2.5*Comparison of Feedback Characteristics Across Teacher, Peer, and AI Feedback*

Characteristics	Teacher Feedback	Peer Feedback	AI Feedback
Expertise & Authority	Positive: Widely valued due to teacher expertise and competence (Hyland, 2006; Yang et al., 2006) Negative: Feedback quality affected by contextual factors and individual teacher variations (Goldstein, 2006; Lee, 2008)	Positive: Provides different perspectives and benefits both receiver and provider (Culver, 2023; Patchan & Schunn, 2015) Negative: Students generally prefer teacher feedback over peer feedback (Zhao, 2010)	Positive: Newer Gen-AI platforms can handle open-ended tasks and higher-order aspects (Dai et al., 2023; Katz et al., 2023) Negative: Lacks deep subject knowledge and contextual awareness (Guo & Wang, 2023)
Timeliness & Workload	Negative: Delayed due to large class sizes and teacher workload (Brown, 2007; Er, Dimitriadis, & Gašević, 2021)	Positive: Helps reduce teacher workload (Chen, 2021) Negative: Requires training for effective implementation (Chen, 2021)	Positive: Provides immediate feedback without workload constraints (Dai et al., 2023; Katz et al., 2023)
Personalization & Context	Positive: Can be tailored to individual students (Brown, 2007) Negative: Varies based on teachers' emotional states and time constraints (Brown, 2007; Yu et al., 2021)	Negative: Less personalized and sometimes affected by friendship dynamics (Loretto et al., 2016)	Positive: Can generate specific, detailed feedback (Guo & Wang, 2023) Negative: Lacks contextual awareness of student abilities and course content (AlGhamdi, 2024; Guo & Wang, 2023)
Dialogic Quality	Negative: Often lacks dialogue due to time constraints; students revise without understanding why (Hyland, 1998; Lee, 2007)	Mixed: While potentially interactive, no specific evidence in the texts addressing peer feedback's dialogic quality	Positive: Gen-AI platforms provide communicative interface unlike earlier AWE tools (Dai et al., 2023; Katz et al., 2023) Negative: May be too lengthy or complex for students to process (Guo & Wang, 2023)
Consistency & Reliability	Negative: Inconsistent due to human factors like mood, fatigue, and subjective judgments (Brown, 2007; Lee, 2008)	Negative: Inconsistent due to multiple reviewers with conflicting suggestions (Loretto et al., 2016)	Positive: More consistent and unbiased as it's not influenced by personal beliefs (Zhu et al., 2024) Negative: May provide unnecessary positive feedback (Dai et al., 2023)

Table 2.5*Comparison of Feedback Characteristics Across Teacher, Peer, and AI Feedback*

Characteristics	Teacher Feedback	Peer Feedback	AI Feedback
Feedback Focus	Mixed: Varies by teacher; some focus on grammar, others more comprehensive (Lee, 2008)	Positive: Effective for rule-based errors Negative: Less effective for non-rule-based errors like word choice (Diab, 2010)	Positive: Can address both lower-level errors and higher-order aspects including content and organization (Dai et al., 2023) Negative: Early AI tools primarily focused on grammatical accuracy rather than content (Cheng & Cheng, 2008; Thi & Nikolov, 2022)
Self-Regulation Support	Positive: Can encourage self-regulation when well-implemented (Hyland, 1998; Lee, 2007)	Positive: Evaluating others' work improves one's own performance (Culver, 2023)	Mixed: Provides process-level feedback with strategies for improvement Negative: Limited personal-level feedback to help with self-regulation and self-correction (Dai et al., 2023)
Implementation Challenges	Negative: Affected by workload, class size, and time constraints (Brown, 2007; Goldstein, 2006)	Negative: Requires student training and teacher intervention for high quality (Chen, 2021)	Negative: Requires higher level of digital literacy for effective prompting and use (Naamati-Schneider & Alt, 2024; Yan, 2023) Negative: Can lead to over-reliance if not properly integrated (Darvishi et al., 2024)
Readability & Structure	Mixed: Quality varies depending on teacher's time and workload (Brown, 2007)	Negative: May lack detail and clarity (Loretto et al., 2016)	Positive: Often more readable than instructor feedback with better fluency and coherence (Dai et al., 2023) Negative: Sometimes too lengthy or complex (Guo & Wang, 2023)
Linguistic Accuracy	Mixed: Generally accurate but varies with teacher proficiency (Brown, 2007; Lee, 2008)	Mixed: Effectiveness depends on peer's language proficiency (Diab, 2010)	Positive: Can make feedback linguistically accurate and contextually relevant (Guo & Wang, 2023)
Best Use Case	Positive: can provide content-focused, personalized feedback that requires subject expertise (Hyland, 2006; Yang et al., 2006) Negative: Limited by time constraints and workload (Er, Dimitriadis, & Gašević, 2021)	Positive: Valuable for developing students' evaluative skills (Patchan & Schunn, 2015; Culver, 2023) Positive: Beneficial when combined with teacher oversight (Chen, 2021)	Positive Most effective when combined with teacher feedback to address higher-order concerns (Wilson & Czik, 2016); excellent for timely, consistent feedback at scale

2.4.3.1 Teacher feedback

Among feedback modes, teacher feedback is widely valued and accepted due to students' trust in teachers' expertise and competence in the subject (Hyland, 2006; Yang, Badger, & Yu, 2006). The literature involves many studies supporting this view. For instance, echoing this finding, Yang, Badger, and Yu (2006) compared two groups of students (one receiving feedback from a teacher and the other from their peers), and found that students preferred teacher feedback to improve their writing. The study suggested that this might be because of students' affective preference for teacher feedback, as the teacher is seen as an authority, whereas peer feedback is more associated with autonomy. This finding demonstrates that teacher expertise is the key factor behind this widespread recognition of teacher feedback.

However, despite this trust towards teacher feedback, the literature also included some studies showing inconsistency of traditional teacher feedback owing to various factors, including contextual factors, as well as individual teacher and student factor. Goldstein (2006) for instance showed that teachers' hectic schedules and the large number of students in classes that create intense pressure on teachers' shoulder which affects their feedback negatively.

Under time pressure, some teachers can also provide varying feedback, which leads to inconsistency in the amount, degree, and quality of feedback. To illustrate, a study conducted by Lee (2008) examined student reactions to teacher feedback in two Hong Kong secondary classrooms by comparing the feedback provided by two teachers on students' writing. This study employed triangulated data to explore the quality of teacher feedback from two teachers by using varied data collection measures including student questionnaires, checklists, and protocols, alongside teacher interviews, classroom observations, and feedback analysis. The findings of this study revealed that student responses to feedback were largely affected by the teacher providing it. The feedback from Teachers A and B varied in focus and provided different amounts of feedback, one being very comprehensive and grammar-focused, while Teacher B focused almost exclusively on grammar, with little to no comments on content and organization. Additionally, Teacher A's feedback was teacher-dominated, as it primarily consisted of direct corrections without requiring students to analyze their errors or self-correct, apart from providing them with coded error feedback.

The literature further highlights another major shortcoming of teacher feedback: its inability to provide timely and personalized responses to each student due to large class sizes and increased teacher workload. In the study of Er, Dimitriadis, and Gašević, (2021), the researchers suggested that although dialogic feedback between student and teacher is necessary, it is not always feasible due to time constraints. In a study by Hyland (1998), this lack of dialogue was seen as a serious handicap for feedback uptake. In his study, where data was collected through retrospective interviews on the use of teacher feedback in ESL writing, it was demonstrated that students revised their drafts without actually knowing why they were doing so due to the lack of dialogue. Similarly, in a study involving interviews with six teachers and eighteen secondary school students in Hong Kong, Lee (2007) also indicated that students revised their drafts without fully understanding the mistakes identified in teacher feedback, which in turn caused them to repeat their mistakes in subsequent drafts.

Another noteworthy limitation discussed in the literature is that teachers' personal views and emotions may influence the feedback process and quality negatively. Brown (2007) noted that providing feedback to large classes can result in inconsistency and delays, as teachers, unlike machines, experience mood swings, fatigue, and subjective judgments, which may lead to variations in the feedback given to different students. Likewise, Yu et al. (2021) also explored teachers' emotional experiences when providing feedback, and findings suggested that feedback is an emotionally charged process for teachers. In the study, it was shown that teachers did not feel positive about providing comprehensive feedback because they were not satisfied with students' responses to their feedback despite their effort. The findings of this study suggested that teachers' attitudes toward feedback can also be influenced by students' responses to their feedback. When they notice that it is not well-received or not used to change things in the paper, teachers may begin to feel emotionally discouraged or demotivated to provide feedback to students' papers, considering their effort pointless.

2.4.3.2 Peer feedback

As an alternative to teacher feedback, peer feedback has emerged as a strategy to reduce teacher workload. However, the benefits of peer feedback extend beyond simply alleviating the teacher's burden. In fact, peer feedback not only supports those who receive it, but also improves the writing skills of the students who provide it. Engaging

in this task allows students to evaluate various components of a peer's work and reflect on potential solutions. This reflective process benefits student assessors, enabling them to apply the insights gained from their evaluations to their own future writing (Patchan & Schunn, 2015). In this vein, peer feedback can be considered a very strong alternative to teacher feedback.

In support of this view, Culver (2023) compared the essay performance of two groups: Group 1 gave and received feedback (traditional feedback), while Group 2 assessed their peers but did not receive any feedback themselves. The results indicated that evaluating others' work is just as important as receiving feedback. Students who assessed their peers improved their performance as much as those who received feedback. However, despite these gains, many students in the non-feedback group failed to recognize the value of the peer assessment process. This finding suggests that the value of peer feedback is not something to be underestimated and is valuable for self-regulated learning.

Another study by Diab (2010) investigated the impact of peer editing versus self-editing on students' ability to correct specific language errors in writing. Using a quasi-experimental pre-test/post-test design, the study compared two groups: one engaged in peer editing (experimental), and the other in self-editing (comparison). It focused on two rule-based errors (subject-verb agreement, pronoun agreement) and two non-rule-based errors (wrong word choice, awkward sentence structure). The results showed that peer editing significantly helped reduce rule-based errors but had no significant effect on non-rule-based errors. This finding suggests that peer feedback was more effective in addressing rule-based errors, where clear grammatical rules can be applied, whereas its limited effectiveness in word choice issues might be related to how skilled the peers are.

In their study, Loretto, DeMartino, and Godley (2016) showed that most students found peer review beneficial for writing development, particularly due to three reasons: anonymity, reviewing peers' work, and receiving feedback from multiple peers. The anonymity of feedback allowed for the negation of friendship pressure. However, the same study also pointed out that despite some benefits of multiple reviewers, it can also create inconsistency in feedback due to conflicting suggestions. For this reason, it is suggested that peer feedback requires skillful evaluation, and when training is not done, the quality of peer feedback may not be sufficient. Therefore, some studies also suggest a peer and teacher combination in feedback or teacher scrutiny. For instance, in a study

conducted by Chen (2021), peer feedback under teacher intervention was examined, and it was suggested that it can be especially beneficial in large classes. Teacher involvement was shown to improve the quality of peer feedback while still helping to reduce teacher workload.

Despite these studies supporting peer feedback, it was still seen that students tend to prefer teacher feedback over peer feedback. Yet, this preference may not reflect actual benefit. For instance, Zhao (2010) noted that students preferred teacher feedback not because it was more effective, but because of perceived teacher authority. In this study, eighteen Chinese university-level English learners were interviewed to learn whether they used teacher or peer feedback more. Content analysis of the interviews revealed that students made greater use of teacher feedback compared to peer feedback, but they often applied teacher feedback to their drafts without fully understanding it. This highlights the need for comparative studies to examine not only how much students use teacher or peer feedback, but also how well they understand and benefit from them. Based on these findings, it cannot be concluded that teacher feedback is inherently more useful than peer feedback. Instead, this preference could be interpreted as an outcome of a teacher-dominated classroom culture, leading to students' blind acceptance of teacher feedback without considering its actual usefulness or their benefit from it.

To conclude, despite its clear benefits, peer feedback has not been widely practiced in classrooms, mainly because students do not accept peers as experts, friendships can influence evaluations, and reviews can be inconsistent due to multiple peer assessors. Additionally, because the teacher is seen as an authority figure in the classroom, students tend to favor teacher feedback over other forms, even when its actual benefit is questionable. For this reason, the literature has recommended peer training or teacher involvement to improve the effectiveness and credibility of peer feedback.

2.4.3.3 AI feedback

As a remedy and response to the shortcomings of teacher and peer feedback, AI-based tools initially emerged as proofreading aids. However, these tools have not been widely adopted in classroom settings, as the feedback generated by Automated Writing Evaluation (AWE) or AWCF tools presents certain limitations for effective classroom use.

Firstly, these tools can mainly resolve issues related to lower level, surface errors (Thi & Nikolov, 2022). Secondly, these AWE tools have no communicative features providing meaningful opportunity to interact with the user (Cheng & Cheng, 2008). Hence, they can offer explicitly corrected version, but do not include comments or any interaction with user that foster students' understanding of their errors by providing error descriptions for learners (Guo et al., 2022; Ranalli, 2018). And as for content-related feedback, AWE tools cannot offer detailed feedback, which made teacher integration of content feedback obligation when these tools were used (Grimes & Warschauer, 2010).

Having seen the limitations of these tools, literature supported the combination of teacher and AI-generated feedback, which was seen to have positive outcomes. For example, Wilson and Czik (2016) conducted an experiment comparing two types of teacher feedback: teacher-only feedback and teacher feedback combined with an AWE tool. The latter combined approach enabled teachers to focus more on higher-order skills, that is, content and organization. AI-based tools are thus believed to reduce teacher workload.

Later, a new generation of AI tools emerged called Gen-AI chatbots with feature of interacting with users that surpass the limitations of earlier AI-based feedback tools. These new emerging platforms, unlike the predecessor AI tools, are endowed with a communicative interface which can handle open-ended tasks and provide feedback for the higher-order aspects, including content and organization of a written task. To address the lack of a human communicative touch, Gen-AI platforms, as AWCF providers, have emerged to elevate AI-based feedback to a more human-like level for open-ended tasks (Dai et al., 2023; Katz et al., 2023).

In Dai et al. (2023), the researchers analyzed feedback on students' research proposals with five rubric aspects: Goal, Topic, Benefit, Novelty, and Clarity, and labelled ChatGPT's and the instructor's feedback on each aspect as positive, negative, or none. These aspects were proposed by Hattie and Timperley (2007) to describe effective feedback. A key finding emerged from this comparative analysis is, the readability of GPT's feedback was better than that of the instructor, with higher scores for fluency and coherence. Regarding the topic aspect, the feedback from the teacher and GPT largely agreed. However, in other areas such as Goal and Benefit, GPT provided more positive feedback than necessary, showing inconsistency with the instructors. Unlike GPT, teacher in this study refrained from offering positive feedback unnecessarily to avoid guiding

students incorrectly. Additionally, GPT provided no personal-level feedback that could help students with self-regulation and self-correction, whereas such feedback appeared in a few cases in the instructor's comments. On the positive side, surprisingly, GPT offered some process-level feedback, providing students with strategies to improve their writing organization. In terms of task-level feedback, both GPT and the instructor were successful in providing feedback on whether the student understand and complete the task properly. The findings highlighted that GPT can provide feedback to open-ended tasks, and yet there is a need for improving some unnecessarily provided feedback, as well as offering personalized feedback to encourage students' self- reflection and self-correction for their writing errors.

It has also been discussed in the literature that, by using AI feedback as a supportive tool, teacher feedback can be rendered linguistically more accurate and contextually relevant than before, thereby decreasing the linguistic burden on educators. In their study, Guo and Wang (2023) showcased this finding. Five teachers reviewed GPT-generated feedback for 50 argumentative essays written by Chinese undergraduate students, focusing on content, organization, and language. Data was obtained via Likert scales and thematic analysis. Unlike traditional AI-based tools, teachers' perception of AI-generated feedback included both positive and negative aspects. To begin with positives, teachers found GPT successful in terms of generation of specific and detailed feedback that included reasoning and a constructive tone providing praise first and then suggestions for student correction. As for the teachers' negative views, they perceived the GPT feedback as sometimes too lengthy, complex, or off focus. Whereas they concluded, students might have difficulty in locating the errors, as GPT did not offer in-text annotation. The feedback also lacked contextual awareness, as it did not incorporate knowledge of students' abilities or course content, and there were a few instances where the feedback was irrelevant or not very actionable.

Zhu et al., (2024) also explored how ChatGPT compares to traditional EFL teacher feedback, and in congruent with study of Guo and Wang (2023), they concluded teacher feedback provided more personalized, context-aware feedback, as the feedback included exam-focused feedback to students. But positively, the researchers stated that feedback provided by AI can be more consistent and unbiased because it is not influenced by any

personal beliefs and pre-formed ideas towards students, unlike teacher feedback which may be subject to personal biases or prejudices.

Similar to the results mentioned above, AlGhamdi (2024) conducted a qualitative study exploring the perceptions of computing students at a Saudi university regarding GPT-generated feedback on their writing skills. The study followed a shift in feedback from teacher to AI without informing students in order to avoid pre-judgments. In this scenario, students received teacher feedback during weeks 1 to 3, GPT feedback in weeks 4 and 5 (without being informed of the switch), and reflected on their feedback experience in week 6. The findings highlighted important insights into both feedback types. Students reported that GPT feedback had a formal and neutral tone, in contrast to the empathetic and encouraging tone of teacher feedback. Students also expressed mixed emotional responses, noting both negative and positive effects of GPT-generated feedback. Furthermore, GPT's suggestions were seen as less personalized, while teacher feedback, though more tailored to students' needs, was perceived as less consistent across students. On the other hand, students often described GPT feedback as clear, consistent, and well-structured, whereas teacher feedback was sometimes considered brief or lacking in clarity.

Darvishi et al. (2024) conducted a study with 1,625 undergraduate students and found that students often became overly reliant on AI tools when providing peer feedback. The findings suggest that while AI assistance helped maintain feedback quality, students did not always internalize the guidance or develop self-regulated strategies for evaluating others' work independently. This finding suggests that students do not know how to benefit from this tool effectively as a learning opportunity.

Furthermore, echoing this view, Yan (2023) suggested that the digital literacy of both students and teachers is insufficient. However, AI literacy is a prerequisite for the appropriate and responsible use of these platforms. As Naamati-Schneider and Alt (2024) reported, Gen-AI platforms require a higher level of digital literacy for effective prompting and fine-tuning of outputs. In their study, students made use of ChatGPT in problem-based learning, and the findings suggested that using AI effectively requires strong AI literacy. This includes knowledge of strategies for double-checking AI responses for accuracy and learning how to prompt effectively to obtain more accurate results.

Derived from culmination of these studies, the common limitations of AI-generated written corrective feedback and possible teacher complementation's for AI Feedback are outlined in Table 2.6.

Table 2.6

AI Feedback Limitations and Possible Teacher Complementation's for AI Feedback

AI Feedback Limitations	Possible Teacher Complementation
Lack of Expertise	Expertise Insights Provided by Teacher
Ethical And Privacy issues	Ethical And Privacy Oversight from Teacher
Nuanced Contextual Understanding	Nuanced Contextual Understanding Strengthened by Teacher Involvement
Constructive and guiding tone	Constructive and guiding Tone Enhanced by Teacher Involvement
Trust And Credibility	Trust and Credibility Increased by Teacher Involvement

2.5 Overview of Existing Empirical Studies on AI feedback

The existing empirical literature showed that the number of studies related to AI feedback is still quite limited and insufficient to provide a complete picture of how AI performs in delivering feedback. In different studies, AI's performance varied depending on the type of task, the training it received, and the nature of student papers because all of these variables can influence the results.

Despite these limitations, the empirical findings are still valuable in that they collectively acknowledge the need for teacher intervention to monitor and refine AI-generated feedback. To systematically present the scope and gaps in the literature, this review categorizes existing empirical studies into three major thematic areas:

1. The growing interest in and acceptance of AI feedback among students and educators and their dispositions towards using it
2. Comparative analyses of AI feedback versus teacher feedback (error and content feedback).
3. Comparisons of different generative AI tools for feedback provision from teachers' perspectives.

Table 2.7 below presents a synthesized summary of these studies, grouped into four thematic categories. This categorization in the table outlines their key findings and, identify areas that require deeper exploration, and demonstrate how the present study approaches the topic differently.

Table 2.7

Overview of Key Empirical Studies on AI Feedback and the Present Study's Contribution

Focus Area	Studies	Key Findings
Dispositions towards AI feedback	Escalante et al. (2023); Henderson et al. (2025); Marzuki et al. (2023); Kurt & Kurt (2024); Jansen et al. (2024); Shoufan (2023);Tossell et al. (2024)	For both parts (teachers and students),AI feedback was seen as accessible, flexible, and useful for idea support, but concerns remained about its inconsistency, lack of contextual depth, overreliance risks, and limited trust in its accuracy and actionability.
2.AI vs Teacher Feedback (Error-focused)	Algaraady & Mahyoob (2023); Lin & Crosthwaite (2024); Yoon et al. (2023) Şahin & Gürman Kahraman (2025)	AI detected grammar issues well, but missed context- related and global issues related to meaning. Teacher feedback was more accurate and actionable. Human feedback can address grammar issues more comprehensively, considering their impact on sentence structure, clarity, and overall coherence, whereas AI's grammar check tends to remain at a more surface level.
3. AI vs Teacher Feedback (Content-focused)	Dai et al. (2023); Farrokhnia et al. (2023); Ray (2023); Steiss et al. (2024); Ziqi et al. (2024) Keser and Pektaş (2024)	AI feedback was readable but vague. It lacked prioritization and guiding tone. Teachers provided clearer, structured feedback.
4. Comparisons Among Gen-AI Tools	Obaidoon & Wei (2024) Dai et al. (2024)	Claude was most precise; GPT-4 nuanced but less actionable; Gemini repetitive and error-prone. GPT-4 was most readable and effective feedback provider compared to GPT 3.5. and human instructors.

2.5.1 Studies on the Perceptions of AI-Generated Feedback

In recent years, an increasing number of studies have explored the potential of AI feedback in academic writing, focusing on either teachers' or students' perceptions. Most

of these studies took qualitative approach and mainly provided insights into the usefulness, clarity, and accuracy of AI-generated feedback, or compared AI feedback with human and peer feedback, and thus provided comparative insights.

Marzuki et al. (2023) conducted a qualitative case study with four EFL teachers to examine the impact of AI writing tools, such as Quillbot, WordTune, Jenni, Paperpal, Copy.ai, and Essay Writer, as well as GPT, on students' writing. This study focused solely on teachers' perspectives, particularly those who were actively employing these platforms. The finding of this study is valuable in that it suggested that teachers unanimously found AI tools beneficial for content and organizational aspects of writing. However, this study also acknowledged a potential bias in its own results stating that this perception may stem from teachers' positive inclination toward these tools since they were active users of these tools. Despite recognizing these benefits, teachers also expressed concerns that long-term use could lead students to develop an overdependence on AI writing tools.

A study conducted by Kurt and Kurt (2024) explored Turkish pre-service English teachers' views on the usefulness of ChatGPT version 3 generated feedback compared to two other sources: peer and instructor feedback. The study offered comparative insights into these feedback sources and employed focus group interviews along with thematic analysis to gather qualitative data. The results indicated that teachers perceived ChatGPT-generated feedback positively, primarily because it allowed them to receive feedback at their own convenience. It enabled users to get responses to their questions or prompts through an interactive dialogue with ChatGPT. Another reported advantage was adaptability: depending on how prompts were written, users could adjust the responses to better meet their individual needs.

However, the same study also identified several limitations of ChatGPT version 3 generated feedback. Specifically, it demonstrated some inconsistencies even when the same prompt was used. Oftentimes, it failed to provide corrections, while in other instances, using the same prompt, it provided proper corrections. The second issue with ChatGPT, highlighted by this study, was that its performance depended heavily on the user's skill in writing prompts. As participants noted, if users do not know how to phrase prompts effectively, they may waste a significant amount of time before receiving the desired response. Another drawback expressed by participants was the lack of a human

tone and personal touch in ChatGPT feedback. Lastly, participants highlighted concerns about overdependence on these tools, as students may rely on them for revisions without incorporating their own voice, creativity, and critical thinking skills.

In terms of comparison between peer, instructor and ChatGPT version 3 feedback, the study also revealed several other noteworthy findings. Accordingly, pre-service English teachers perceived instructor feedback as the most reliable, accurate, relevant, and in-depth among the three sources they compared. Students trusted teacher feedback without questioning its accuracy, whereas both peer and AI-generated feedback were sometimes inconsistent, and therefore they were not found reliable as that of instructors' feedback. In terms of relevance, teacher feedback was also perceived as more responsive to student needs, and likewise in terms of depth, teachers could pinpoint minor issues that ChatGPT and peers often failed to notice. However, for content development suggestions, both ChatGPT and teachers were preferred over peer feedback. In short, this study concluded that ChatGPT could be useful tool for the feedback process as a supplement tool because it was not found as effective as teacher feedback, but nevertheless, AI-generated content feedback was found more reliable and useful from teachers' perspectives.

Another line of research has focused on student perceptions of AI feedback. For instance, Escalante et al. (2023) investigated students' perspectives on feedback received from GPT-4o and human instructors to examine student preferences and the impact of AI feedback on students' writing quality. Their study focused on six key areas in feedback: topic sentence quality, idea development, academic language, use of transitional phrases, integration of sources/evidence, and grammatical accuracy. In the output, AI provided grammar feedback in a table format and content-related comments under the targeted headings mentioned above. The results showed that both AI and human feedback contributed to improvements in students' academic writing over a six-week period. Notably, students expressed an equal preference for both feedback sources, highlighting the growing potential of AI tools in writing instruction.

However, in the same study, students wrote 300-word paragraphs that integrated sources from readings. This might have impacted the outcomes of AI feedback, as the students wrote using sources affecting their writing content and organization for good. It is unfortunate that the appendix of this article included only one sample paragraph, which limits insight into the language proficiency of other students and the nature of their errors

across the dataset. Based on the sample paragraph included in the appendix, it can be concluded that the overall meaning of the text was generally comprehensible, with no major meaning level breakdowns. This suggests that, in this particular study, the types of errors might have positively influenced AI's performance. Therefore, a key limitation of this study is the grammatical errors identified by AI did not involve meaning-related issues, and plus, because of the task type, which allows students to draw on external sources for their writing, the students' writing may not include dense errors. As a result, this study provides only a limited perspective on AI's effectiveness with struggling writers, whose meaning-related problems often stem from loose and problematic grammar. This, in return, might have affected students' perceptions related to AI feedback in a positive way.

In the study by Shoufan (2023), students' perceptions of GenAI feedback were explored using a mixed-methods design. A total of 248 university students were asked to use ChatGPT (GPT-3) and then write reflective comments on their experiences. Following a qualitative analysis of students' open-ended reflections, a survey questions formed based on the qualitative analysis was conducted to validate the qualitative results. The study focused specifically on perceptions of GPT-3, and the findings suggested that while students expressed concerns about the accuracy of AI feedback and called for a mechanism to fact-check and verify its responses, they still recognized ChatGPT version 3 as a valuable tool for learning, idea generation, and professional development..

In their study, Tossell et al. (2024) examined undergraduate students' perceptions of ChatGPT's feedback in the writing process and their views on AI-based grading. The version of ChatGPT used was not specified, but it was likely based on the GPT-3 family. Researchers gathered students' perspectives by conducting questionnaires before and after students completed essays with ChatGPT's assistance. The findings indicated that while students appreciated using AI feedback as a learning opportunity, they also felt that it did too much of the thinking for them, which could lead to learning loss. As a result, they expressed a need for teacher support and guidance on how to use the tool effectively. Students also reported that they did not fully trust AI feedback and questioned the accuracy of its output. Furthermore, they stated that they were hesitant about being graded solely by AI without teacher involvement. Overall, they preferred teacher grading, and

many expressed a preference for a hybrid approach that combines AI-generated and teacher feedback, or a feedback crosschecked and validated by teacher.

Jansen et al. (2024) further explored perceptions towards AI feedback by conducting a qualitative study on student teachers' perceptions, who can provide insights from both perspectives. To eliminate bias in results, participants evaluated AI-generated feedback without knowing whether it came from AI or a human. This study utilized from Hattie & Timperley's (2007) feedback model, which addresses three key aspects of student learning: "*Where am I going?*", "*How am I going?*", and "*What's next?*" (guidance for improvement). GPT-3.5 Turbo feedback was evaluated using three scale items: "*This feedback is useful,*" "*This feedback is helpful,*" and "*This feedback provides a lot of support.*" The findings revealed that AI feedback has yet to match the usefulness of teacher feedback. However, AI feedback was accepted in 59% of cases and preferred over teacher feedback in 23% of direct comparisons. While AI feedback shows promise, it remains less effective than teacher feedback due to its higher lexical variance, which negatively impacts actionability of feedback.

In a very recent study, Henderson et al. (2025) conducted a large-scale cross-sectional survey to investigate how university students who use and do not use AI perceive the usefulness and value of AI-generated feedback compared to teacher feedback. The findings from the survey compared two groups of students' perceptions. Students who already use AI for their assignments appreciated GenAI for its easy access and availability. In terms of quality, they reported that GenAI was less critical of their work, which made them feel less judged. They also found the feedback more understandable. However, they criticized AI feedback for being less actionable and lacking context ;often providing vague and generic advice, and missing important details that a teacher would typically catch.

Overall, these findings suggest that AI feedback is valued, and yet the suggested use is, as highlighted, under human supervision to mitigate student concerns and improve reliability.

2.5.2 Comparative studies examining AI and human feedback on student writing

A growing body of research has also conducted comparative studies that examined AI-generated feedback in comparison to human feedback in with a focus on its performance of error-based and content-based (textual) feedback. However, the findings of these

studies are also based on the pre-trained AI platforms, that is, they were not specifically fine-tuned for classroom feedback, as also confirmed in the study of Xiao et al., (2024).

To start with, Yoon et al. (2023) studied the performance of GPT-3.5 in comparison to teachers in evaluating coherence and cohesion in 50 selected argumentative essays written by ELL students. Their findings indicated that while GPT could identify the presence of connectives within sentences, it struggled to assess whether these connectives were used appropriately. Furthermore, GPT frequently overlooked major content issues owing to surface-level analysis, such as repetitive ideas at the paragraph level, while it could only provide feedback on repetition at the word level. This suggested that discourse or organizational-level repetitions remained unaddressed by GPT. The study also indicated that GPT focused primarily on grammar and vocabulary yet missed unclarity in meaning or sentences disconnected from the context. The researchers also implicated in their study that this situation can potentially lead to essays with surface-level grammar issues being heavily penalized with low scores, while essays containing deeper semantic problems might escape proper evaluation and end up with unjustified high scores. The findings of the study also showed that GPT alone was not sufficient to provide students with actionable feedback for effective revision, as it tended to be too generic and lacked the specificity needed to guide students constructively.

Focusing on error-based feedback, a recent study conducted by Lin and Crosthwaite (2024), GPT-4 and human WCF feedback was compared, and the feedback practice of teachers and GPT-4 were found to respond noticeably different. Teachers provided mix of direct and indirect written corrective feedback (WCF) to local and global errors, while GPT provided feedback more in the form of reformulation and metalinguistic feedback. This study revealed both feedback from teachers and GPT-4 had different shortcomings due to different factors. While teachers were sometimes inconsistent in their feedback practice due to tight deadlines and other human factors, GPT-4 failed to grasp contextual understanding in the sentence uttered, and it performed corrections unnecessarily for accurate elements.

Algaraady and Mahyoob (2023) also conducted a study comparing the performance of teachers and GPT in detecting surface- and global level grammar errors in EFL learners' essays. Their findings indicated that although ChatGPT can identify surface-level errors, it failed to recognize issues related to meaning, structure, or the pragmatics of writing. In

contrast, teachers easily noticed these meaning-level issues and provided feedback on them. Their study also concluded that GPT could be useful for providing feedback but should not be considered a substitute for teacher feedback.

Different from these, Lu et al. (2024) compared ChatGPT and human feedback in terms of quality and quantity, using a mixed-methods design. They collected data through quantitative analyses, including counts of feedback types, and qualitative data such as student interviews exploring perceptions and the quality of feedback. The study found that ChatGPT tended to provide general suggestions, while teacher feedback offered more specific guidance for improving writing. Teacher comments addressed various aspects of student writing, such as grammar, organization, and coherence. ChatGPT provided fewer comments, but these were still considered meaningful for content improvement. Ultimately, owing to the clarity and context-specific nature of teacher feedback, students perceived it as more specific and actionable compared to GPT feedback.

Guo and Wang (2024) compared teacher-provided feedback with GPT-generated feedback in terms of the comprehensiveness of comments on Chinese EFL students' argumentative essays. This study highlighted that GPT can provide more detailed and comprehensive feedback compared to teachers across all areas of writing, including content, organization, and language. The overall findings also suggested that GPT can enhance teacher feedback by offering more comprehensive insights into error types within a written work when used alongside teacher feedback.

A significant limitation of the study of Guo and Wang (2024), as acknowledged by the researchers of the article, was its lack of evaluation of GPT's performance across various error types. This occurred because the prompt used by the researchers did not involve any training with error codes. The study hence highlighted the need for future research to conduct a more comprehensive analysis of GPT's ability to identify and address different error types when error codes are used in feedback. Additionally, this study also implicated that future studies should explore whether AI generates redundant codes under these conditions and whether overgeneralization occurs with specific error categories.

In another study, Şahin and Gürman Kahraman (2025) compared GPT-4's feedback with teacher feedback on improving B2-level Turkish EFL students' argumentative and opinion essays. Both the teacher and GPT-4 were asked to give direct corrections, placing the corrected version in parentheses immediately after the error. The results showed that

teacher feedback was more comprehensive, covering not only grammar and spelling but also grammar issues that influence sentence structure, clarity, and meaning. By contrast, ChatGPT-4.0 focused mainly on local, surface-level corrections, such as grammar, spelling, and small word choice changes. The authors suggested that teachers and AI should work in collaboration; AI could handle initial surface corrections, while teachers could focus on deeper level grammatical issues that affect structure, meaning, and clarity.

With a focus on content feedback only, several studies carried out. One of these was, Dai et al. (2023), who indicated in their study that AI-generated feedback is similar to teacher “textual” feedback on student proposal reports in aspects of goal, topic, novelty, benefit and clarity. In their study, they employed GPT-4 to provide corrective feedback to undergraduate writings, and their study have reached important findings. Accordingly, GPT can offer readable and more comprehensive feedback compared to teacher, and in high agreement with teacher, almost all aspects of textual feedback. Similarly, another study on GPT also underscored the potential of AI to provide more comprehensive and descriptive “holistic content feedback” to student papers (Farrokhnia et al., 2023; Ray, 2023).

In another recent study conducted by Steiss et al., (2024), the quality of feedback provided by non -pro plan (ChatGPT 3.5) was assessed compared to teacher feedback by recruiting a team of research personal trained to evaluate AI feedback quality on scale of 1 to 5. The results indicated, feedback generated by GPT was promising for the essay genre-based feedback, also called ‘criteria feedback’ whereas teacher feedback was better in clarity, accuracy, prioritization of essential features and supportive tone. This suggested AI fell short in these aspects and should improve these to achieve teacher alike feedback. Similarly, Ziqi, Xinhua, Qi, & Wei, (2024) highlighted the same necessity, stating teachers need to evaluate AI feedback, especially for content-focused feedback. It also emphasized the importance of informing students about the strengths and barriers of feedback on content to guide them in using AI systems effectively and help them acknowledge these limitations while using such tools.

In their study, Keser and Pektaş (2024) compared the quality of formative feedback provided by ChatGPT and experienced human evaluators on academic argumentative essays written by advanced-level Turkish university students. The essays were evaluated on four criteria: alignment with assessment criteria, clarity of improvement suggestions,

accuracy, and supportive tone. The findings showed that human feedback was generally rated higher in three level involving clarity, accuracy, and supportive tone because their feedback were more precise, actionable, and offered contextually relevant guidance. In contrast, AI slightly outperformed humans only in criterion-based evaluation. While supportive tone scores were similar, human feedback retained a slight advantage due to its empathetic and relational nature.

2.5.3 Comparisons among different gen-AI for feedback provision

Although limited in number, the literature includes only one study that compare the performance of different AI models in writing context. These studies have evaluated the accuracy, usefulness, and limitations of different platforms in comparison.

To start with, a study of Obaidoon and Wei (2024) involved the comparison of several non-pro subscription plans of ChatGPT, Gemini, Bing Chat, and Claude. Feedback from these non-pro AI plans was evaluated in the context of Chinese foreign language writing essays. In their study, four educators with different educational backgrounds reviewed four essays written by Chinese students of varying language proficiency. Meanwhile, the AI platforms also reviewed the same set of essays to provide short-written feedback on two aspects: errors and rhetoric, and accordingly, grade student papers. The findings of this study suggested valuable findings thanks to its comparative approach. For instance, it concluded that the non-pro plan of Claude outperformed the other AI platforms in the feedback provided to error and rhetoric. The study also revealed that GPT-4 had a nuanced understanding of grammar, but by comparison to Claude, it could provide less actionable feedback. As for GEMINI, it was more precise and consistent in feedback, but it still failed to recognize grammatical errors and tended to be repetitive.

On the other hand, the same study indicated that Claude showed superior performance with its near-precise holistic feedback on surface-level issues such as grammar and vocabulary, along with specific written comments for content feedback. However, the feedback from Claude somewhat fell short of pinpointing the positive aspects of student writing and guiding corrective feedback or suggestions for improving student writing. These findings suggested that Claude's performance could be significantly enhanced through teacher filters or prompt engineering and thus could be rendered a more reliable tool for educational purposes. However, as also confirmed by the study, the type of feedback studied in this study was not comprehensive one, but holistic in its nature,

consisting only of comment-based suggestions with a few remarks. This highlights the need for further research into the deployment of pro-versions to determine whether they can offer more comprehensive feedback on both error and content, as required in classroom settings.

In another recent study conducted by Dai et al. (2024) investigated how well two versions of GPT models, GPT-3.5 and GPT-4 could generate textual feedback for writing tasks in a data science course, comparing feedback from both AI models to that of human instructors. In terms of readability, it was found out that both GPT-3.5 and GPT4 were able to generate more readable feedback consistently than human instructors, which is noteworthy finding in that it suggests AI's ability to have consistent format, clear language and organization in feedback provision, while newer version GPT4 provided more effective feedback than both GPT 3.5. and human instructors in terms of feeding-up, feeding-forward, process level, and self-regulation level.

2.5.4 Identified gaps and directions for future research

Several specific gaps emerge from this review. Most studies in the literature involve GPT alone, while Claude and Gemini were employed only in a handful of studies. Moreover, in most studies, it was not specified whether AI tools were trained or fine-tuned prior to implementation. It also remained uncertain whether GPT was chosen simply because it was the only available or accessible tool at the time of data collection, which may explain its widespread use in previous studies. However, in either case, the interpretability and generalizability of the findings in these studies remained limited due to the lack of clarity regarding tool selection and absence of training procedures.

A more systematic outlook was required to understand the usability of AI-generated feedback in classroom contexts. These studies highlighted collectively need for teacher filter. Ninaus and Sailer (2022) argued that AI in education should not be fully automated, teachers must remain in control. Hence they proposed a closed-loop model for integrating AI feedback into teaching, which involves the following steps to carry out this: (i) data recording, (ii) pattern detection, and (iii) adaptivity. This framework provides valuable groundwork for the present study, as it supports teacher involvement. Adapting these terms to the present study: data recording may refer to documenting how teachers refine AI feedback, what they change, and why; pattern detection can involve identifying

common refinement needs; and adaptivity refers to using those patterns to adjust or improve how AI feedback is generated in the future.



3 METHODOLOGY

3.1 Research Design

The main aim of this study is to explore how teachers refine AI-generated feedback and evaluate its accuracy based on their refinement decisions, using a qualitative approach. In addition, teachers were interviewed to gain insights into the strengths and limitations of this process and to elicit their suggestions for future classroom use of AI-generated feedback.

3.1.1 Multiple case study design

In congruent with the research phenomenon, this research aims to employ multiple case study approach. According to Yin (2009), multiple case study designs can be categorized as: (a) multiple cases with a holistic design and (b) multiple cases with an embedded design. Among these designs, the holistic multiple case study approach was selected, as this can enable providing answers to research questions in a more comprehensive manner. The rationale for selecting this design is explained in detail below.

The study comprises eight teachers (n=8), with each teacher representing an individual case. The reason for having such a small sample in the present study lies in the purpose of qualitative designs, which aim to explore phenomena in depth rather than to generalize statistically. This resonates with the goals explained by Yin (2009) for multiple case studies, in that the multiple case study design does not aim for statistical generalization but rather for analytical generalization, where the richness and depth of each case contribute to a broader understanding of the phenomenon.

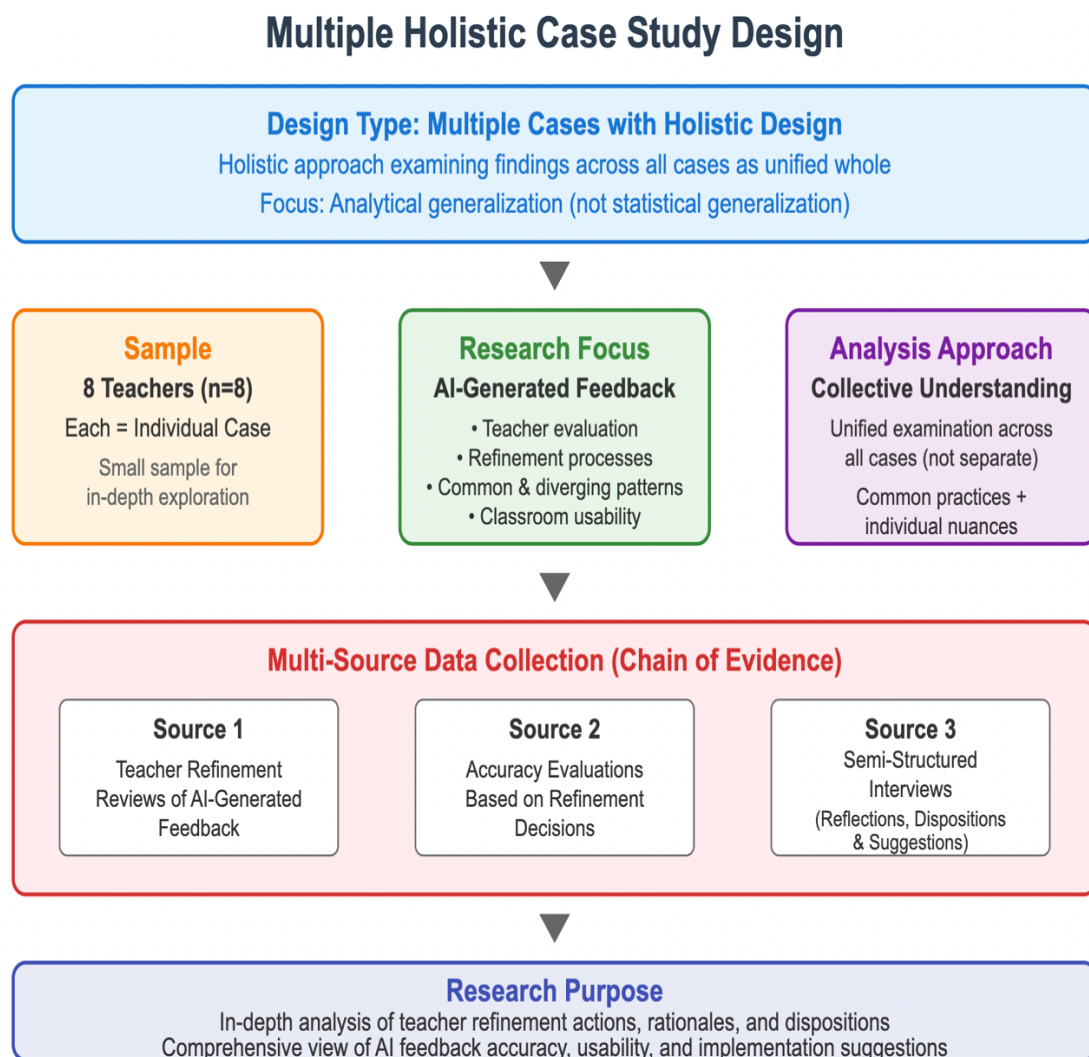
Unlike embedded designs where subcases exist within each primary case and the cases analyzed separately, the multiple holistic case design takes a holistic approach by examining findings across all teacher cases as a unified whole. Thanks to this approach, common and diverging patterns in how teachers evaluated and refined AI-generated feedback were identified. Instead of examining each teacher's process separately, the design emphasizes a collective understanding of how teachers interact with and respond to AI feedback, and thus the study aims to offer comprehensive view of the phenomenon that reflects common practices as well as individual nuances.

In case studies, it is important to collect data from multiple sources for the same purpose in order to establish a chain of evidence (Yin, 2009). For this reason, data were collected

through teacher refinement reviews of AI-generated feedback, accuracy evaluations based on their refinement decisions, and individual semi-structured interviews. This multi-source approach enabled an in-depth analysis of both their refinement actions and their reflections on the accuracy and future classroom usability of AI feedback, as well as the conditions necessary for such use.

Figure 3.1

Overview of Research Design and data sources



3.2 Context of the Study

The present study was conducted at the School of Foreign Languages of a private university in Türkiye, with English instructors as participants. The aim of preparatory program is to equip undergraduate students, who already have a basic understanding of English, with the skills needed to enhance their proficiency in all four language areas for academic and everyday use. Each module lasts for eight weeks and is designed to develop

students' competence in reading, writing, listening, and speaking. At this institution, the four skills are taught as separate components, each allocated its own instructional time.

In the B1-level modules, academic writing classes are allocated six hours per week. The course specifically aims to help students develop the ability to produce well-structured academic essays. By the end of the module, students are expected to clearly express their opinions in structured essays that demonstrate linguistic accuracy, strong content, and a well-organized flow of ideas supported by cohesive devices.

During this level, students are taught two different academic writing genres. They learn how to write both opinion essays and cause-and-effect essays. However, opinion essays serve a pre-academic purpose and are usually limited in length (minimum 250 words) as a starting point for essay writing in the institution's curriculum following the A2 level, where students have only written very short paragraphs and have not yet been exposed to the essay genre. Due to students' novice writing level, feedback for opinion essays aims to help them build basic essay structures with more holistic and explicit guidance. As such, they do not present a sufficiently robust context for the type of analysis conducted in this study.

Cause-and-effect essays, on the other hand, are introduced after the midterm. At this point, students become more independent academic writers and are expected to produce essays of at least 350 words. Accordingly, they are provided with more comprehensive and structured feedback focusing on both error and content to support their self-development as independent writers. For cause-and-effect essays, students receive metalinguistic feedback, including error codes and content-based feedback. Truscott (2007) outlines this approach as a strategy that helps students identify and discover their errors by themselves without direct correction from the teacher. This is why the present study focused only on the students' cause-and-effect essays and the type of feedback provided for them, as this feedback involves more structured feedback that can uncover a variety of pedagogical reasoning strategies employed by teachers when providing feedback. Thus, it can be examined whether the performance of AI-generated feedback is comparable to that of teacher feedback through a rigorous analysis.

Within the institution, this mode of feedback is provided by the teacher only once upon first draft. In this process, students are expected to revise their writing by interpreting the error codes and guiding comments independently. However, this type of feedback, being

comprehensive in its scope, presents several challenges for instructors in preparatory programs. Given the large number of essays reviewed each module and teachers' hectic schedules, instructors must evaluate various aspects of writing, ranging from grammatical accuracy to content quality, within a limited timeframe, often in the remaining time after classes or even on weekends. In this context, a typical writing instructor is responsible for reviewing at least 15 essays per class in a meticulous manner. As a result, balancing error correction and content-focused feedback while ensuring consistency and fairness challenge teachers cognitively.

In response to this ongoing challenge, AI has begun to be seen as a potential support tool. For this reason, pro-plans of three AI models (GPT-4o, Gemini Advanced, and Claude 3.5 Sonnet) were trained to test their capabilities to provide feedback similar to that of teachers. They were trained based on the institution's error codes because these codes were developed specifically for B1-level preparatory class students, offering a simplified and targeted system, and they became familiar with this code system since A2 level. AI models were additionally guided by teacher feedback samples provided by the Testing Office of the institution. These samples included real student essays annotated with both error codes and content feedback from teachers. In this vein, the AI-generated feedback consists of the following three components:

1. **Error-code incorporated student papers** with metalinguistic symbols used at the institution.
2. **Error tables** that include sentences with errors and metalinguistic errors to cater for descriptions which normally teacher would do orally, to enhance teachers' metalinguistic feedback with cues and prompt-type questions.
3. **Content-based feedback** on strengths and areas for improvement, including comments on essay components such as the thesis, supporting ideas, and conclusion. It also uses a constructive and guiding tone with cues and prompt-type questions and aims for student self-correction.

To carry out the evaluation, ten cause-and-effect essays written by B1-level students on two different topics were selected. Among them, one essay was designated as a pilot paper. This pilot allowed teachers to become familiar with the feedback format and review procedures before starting the actual evaluation. Each participant reviewed the

same set of papers, which ensured reviewing and refining same papers under comparable conditions.

3.3 Participants of the Study

This study employs a sample of eight ($N = 8$) preparatory-class English instructors. Both convenience sampling and criterion sampling were used for selection of participants. Convenience sampling was primarily used to select participants based on their availability and willingness to take part in the study (Mackey & Gass, 2022) while criterion sampling, a type of purposeful sampling were used to select participants based on specific criteria that align with the research objectives (Patton, 1990). As a fundamental criterion, familiarity with feedback practice was considered as paramount. Therefore, participants were selected because they were already familiar with the feedback approach used at the institution through their prior experience teaching in preparatory classes and their ongoing participation in institutional standardization protocols for feedback. This experience was crucial because this provided a solid background to teachers for comparing their established feedback standards to those generated by the AI system. On the other side, since the study focused exclusively on the evaluation of AI-generated feedback, rather than its creation, prior experience of feedback with AI tools was not considered essential to participate in this research.

To mention the participant profile briefly, the teachers' profiles varied in terms of teaching experience years, ranging from 2 to 19 years, but this allowed for diverse perspectives in evaluating AI feedback. Only two of the participants had practical experience to test AI's capability to provide feedback for student writing papers though still they noted their experience was still very limited. The rest of the participants did not mention any use for feedback, and stated they only used AI for lesson planning and material development. And most stated their knowledge of AI is only limited to ChatGPT, and they have no familiarity to other platforms of AI.

Table 3.1. provides demographic information about the teachers as well as their prior experience with AI for writing feedback. To ensure confidentiality, their names were anonymized in the given table and represented using the code "T" followed by a number:

Table 3.1*Demographic Information about the Participant-Teachers*

Participant	Age	Gender	Teaching Experience	Educational Background	Qualification	Prior Experience with AI For Writing Feedback
T1	42	Female	19	English Language & Literature	Bachelor	Had never used AI for providing feedback before this study and explored its capabilities during the research.
T2	26	Female	2	English Language Teaching	Bachelor	No prior experience in using AI for feedback; has only used it for lesson planning and material development.
T3	30	Male	8	English Language & Literature	Master	Familiar with AI platforms; used AI for providing feedback and for his research on AI feedback.
T4	26	Female	3	English Language Teaching	Master	Has not used AI for feedback, only for lesson planning and material design.
T5	27	Female	4	English Language Teaching	Master	Used AI to get an idea of how to give feedback on student papers, but not for students' use.
T6	37	Female	17	English Language Teaching	Master	Used GPT once for an essay but found it mechanical and robotic.
T7	29	Female	6	English Language & Literature	Master	Knows GPT and occasionally uses it to check her own feedback on student writing. Does not rely on AI for full feedback due to accuracy concerns.
T8	30	Female	6–7	English Language & Literature	Master	Used GPT to provide feedback for only A1-level students and found it very accurate.

3.3.1 Teacher profiles

Since each teacher brought a unique professional background and approach to feedback, shaped by their teaching experience and educational training, having insight into each teacher's profile is crucial for understanding their approach to feedback. For each teacher, profiles created based on interview data are presented below:

While all participants in the study went through standardization sessions to ensure a certain level of consistency in how AI feedback needs to be refined, such efforts do not eliminate personal variation entirely. Instructors, regardless of training may still differ in how they interpret and respond to AI feedback, even if the overall procedures are similar. Hence, having insight into each teacher's profile is crucial for understanding professional identity better.

T1 holds a BA degree in English Language and Literature. She is the most experienced teacher among the participants, with nearly 9 years of teaching experience across various educational segments, including language courses, colleges, and university-level preparatory classes. Since 2015, she has been working at the same institution. She has taught B1-level students many times during her career and is both comfortable and accustomed to providing writing feedback to the university's student profile, following the common practices and format provided by the institution.

T2 is a graduate of the English Language Teaching department. She has been working at her current institution for two full years, during which she has taught B1-level writing classes multiple times in different modules. As a result, she is familiar with the university's feedback format and student profiles. Regarding her use of AI, she primarily uses it to create lesson plans and develop writing materials and activities for students. Additionally, she utilizes ChatGPT to provide holistic feedback to whole class to highlight students' common errors.

T3 holds both a BA and an MA degree in English Language and Literature. He believes he is skilled in using AI, particularly for creating lesson plans and providing holistic feedback on student writing, and his ongoing research on AI as a PhD student has improved and extended his skills and knowledge of AI. His work experience spans various levels, including university-level teaching, General English at the K-12 level, and one year of IELTS instruction. Of his extensive experience, the last three years have been at the university level. At his current institution, he is about to complete his first year.

During this time, he has taught B1-level students and is, therefore, familiar with the B1-level writing feedback practices of the institution and the requirements of this study.

T4 is a graduate of the English Language Teaching department and is currently pursuing a master's degree while writing her thesis in the same department at a state university. She has three years of experience teaching preparatory-class students at the university where the research is being conducted. She has also taught B1-level students multiple times and is knowledgeable about the institution's feedback format as well as the student profile at the university. She also expressed that she uses AI mostly for creating activities and worksheets for the classes.

T5 is a graduate of the English Language Teaching and holds a master's degree. Additionally, she holds CELTA certificate and is currently undertaking DELTA courses to further enhance her teaching expertise and qualify for tutorship. She has three years of experience teaching preparatory-class students in this university. In the context of the research, she has taught writing classes several times to B1 level students, provided feedback on both errors and the content of student essays, and is familiar with the feedback format, including error coding and content feedback, used within the institution.

T6 holds a BA degree in American Culture and Literature, and currently pursues MA in English Language Teaching. Having obtained CELTA certification and pedagogical formation training as well as 17 years of teaching experience, including 11 years of preparatory-class instruction at her current institution, her contribution to current study is noteworthy. T2 expressed that she frequently uses AI platforms for her lesson plans and provide writing feedback on students for error identification and content-related suggestions. These experiences make her a valuable contributor to the research context.

T7 holds a BA and an MA in English Language and Literature. She is currently in her third year at her present institution, but she has a total of six years of teaching experience. She is familiar with feedback practice performed in this institution very well, as she has also taught B1 level students for writing. She is also a user of AI, but her experience with AI is limited to creating activities and designing lesson plans. She expressed that she did not use AI for writing feedback, as she found the results unreliable.

T8 also holds BA and MA degree in English Language and Literature. Additionally, she holds a CELTA certificate. She has 4 years of experience teaching preparatory-class students, plus three-year general English in language courses. In this institution, she is

about to complete her first year. Like the other participants, she has taught B1-level writing courses and is familiar with providing both error and content-related feedback based on her previous and current teaching experience. Her experience with AI is very limited. Among AI tools, GPT is the only platform she is familiar with, and she has used it once with A2-level students to provide grammar feedback; however, she has not used it for content feedback.

3.4 Training

3.4.1 Training of AI platforms

For selecting the appropriate AI for the scope of this feedback, three advanced pro plan of generative AI platforms available at the time of this study, GPT-4o, Gemini Advanced, and Claude 3.5 Sonnet, were trained to provide teacher-like accurate feedback in error identification and tone. The intended feedback output comprises three key components: an error-incorporated student paper, an error table with detailed error descriptions, and targeted content feedback that mentions strengths and addresses only specific areas for improvement.

The training data was derived from prep-class teachers' feedback samples on B1-level students' essays, which included metalinguistic feedback with the institution's error code symbols, as well as content-related comments of brief notes in a guiding tone. To fine-tune the AI based on the feedback practices implemented at the university, sample teacher feedback was obtained from the Testing Office and employed by the researcher during the training process. By grounding the training data in authentic classroom teacher feedback, the study aimed to capture the expectations of classroom-based teacher feedback implemented in the institution.

Since the performance of AI models heavily depend on the quality of its training data, inaccuracies or biases in the input data can lead to unrealistic as well as misleading results (Daneshjou et al., 2021; Gaskins, 2022). For this reason, the principles outlined by White et al. (2024), which involve providing AI with clear prompts, detailed context, and samples, as well as iterative testing to train AI, were followed to optimize AI to produce more accurate results. Additionally, in line with Kouassi's (2022) key learning approaches, the training process was structured into three main steps:

- *Supervised Learning (SL)*: Using labeled data to guide the AI in generating the desired feedback (sample teacher feedback, error code file and error descriptions with examples).
- *Reinforcement Learning (RL)*: Reinforcing accurate outputs and correcting inaccuracies through continuous evaluation (reinforcements with corrections)
- *Deep Learning (DL)*: Refining feedback through iterative fine-tuning to ensure consistent and comprehensive results (iterative training with sample papers)

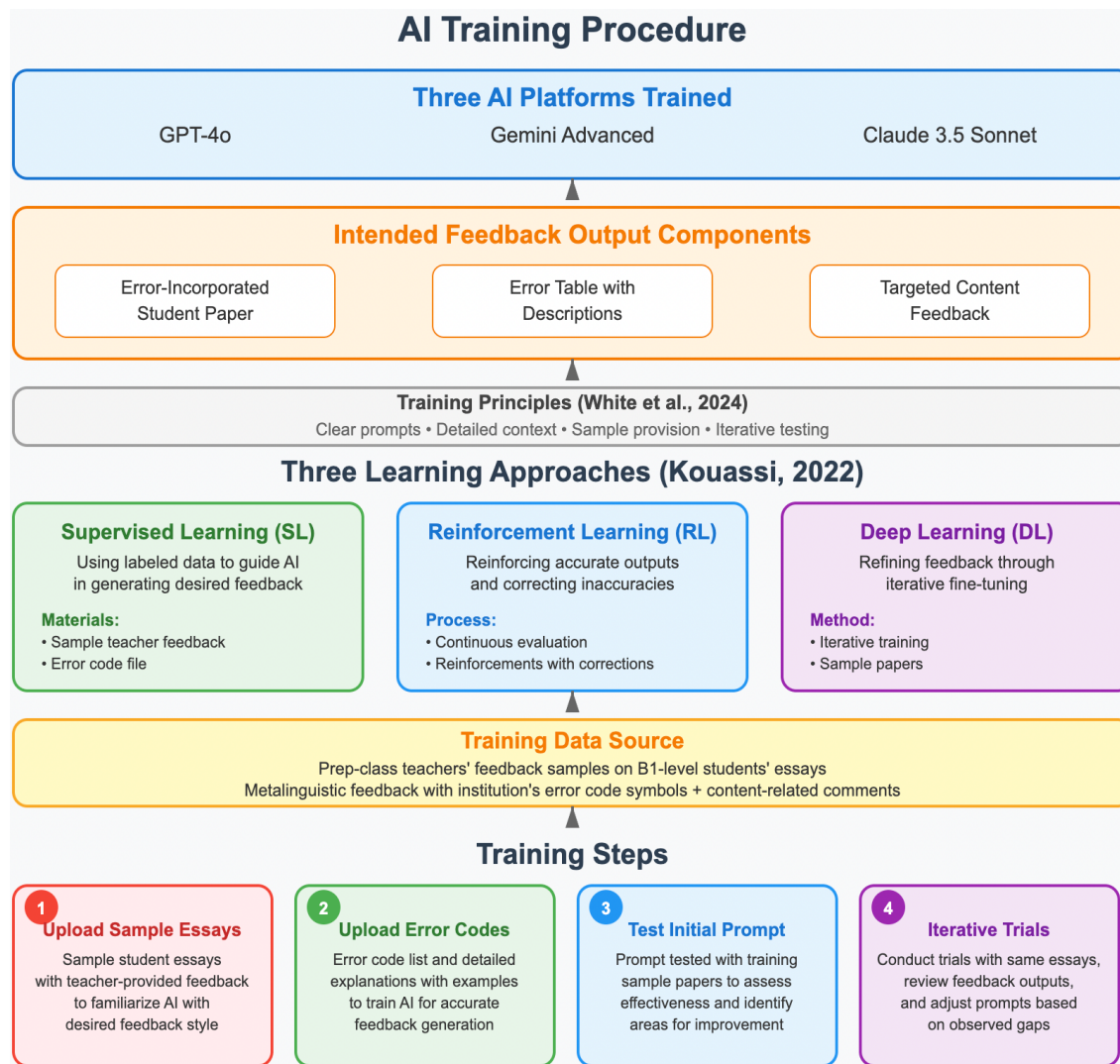
To integrate these three learning approaches into the training of AI platforms, several training materials were employed. These were also used to improve accuracy of AI-generated feedback (See Appendix-1 for the AI training materials). The steps followed for training the AI are outlined below:

- *Uploading Sample Student Essays with teacher-provided feedback*: a subset of seven sample student essays with teacher-provided feedback was uploaded to AI system to familiarize AI with desired feedback.
- *Error Code Files*: Both the error code list and another file where detailed explanations of error codes with examples were uploaded to help train the AI for accurate feedback generation.
- *Testing the Initial Prompt*: The prompt was tested with the training sample papers to assess its effectiveness.
- *Conducting Iterative Trials*: Iterative trials were conducted with the same essays. Feedback outputs were reviewed, and the prompt was adjusted based on observed gaps.

The summary of the procedure followed to train the AI platforms can be seen in Appendix 1, Section 1.4. The steps of the training procedure, including information about the intended feedback, training principles, learning approaches used, and specific steps, are presented as a visual summary in Figure 3.2.

Figure 3.2

Summary of the training steps for the Gen-AI platforms for feedback provision



3.4.2 Selection of Claude 3.5 sonnet as an AI for providing feedback

In the comparative analysis of three pro-plan AI platforms' performances, Claude 3.5 Sonnet demonstrated several advantages for both overall performance and task-specific capabilities. First, it enabled a convenient and user-friendly training process due to its artifact feature, which makes all modifications visible, and all attachments uploaded remain accessible during feedback generation. More importantly, Claude 3.5 Sonnet provided consistent feedback for the same paper in a single call, whereas GPT-4o varied in output due to its high creativity, requiring multiple clarifications to achieve stable output. Gemini Advanced showed the least stability, with responses frequently being cut off mid-feedback and hallucinations occurring during generation, necessitating multiple attempts to obtain complete, reliable feedback.

The reason for this instability in output can be attributed to the token limit of (input + output) capacity of AI model. The information retrieved from OpenAI Tokenizer showed that the prompt employed for this study is ~2,767 tokens, which is well within both models' context limits (Claude 3.5 Sonnet's massive 200,000 compared to GPT-4o's still impressive 128,000). The actual prompt used was about 1,500 tokens, with outputs running around 1,226 tokens, which should have been manageable for their roughly 2,048-token response capacity.

On paper, both GPT-4o and Claude 3.5 Sonnet should handle this workload in a single call, but it was found in practice that GPT-4o started struggling in longer exchanges. Breaking the task into smaller chunks did help GPT-4o produce more reliable output, but its creativity did not allow for stable output for the same paper across different trials.. Gemini-advanced was the most problematic among the three, not only delivering cut-off feedback but also producing hallucinations during the feedback generation process. Claude 3.5 Sonnet, however, seemed to take everything in one call thanks to its 200,000-token context window, significantly larger than GPT-4o's 128,000, and could produce more stable output. In this respect, Claude's ability to keep the context fresh in its memory gave it precedence over the other rivalries, although it was not flawless. The technical capabilities of AI models can be seen in Table 3.2. comparison to each other:

Table 3.2

A summary of Technical Capabilities of AI models

Technical Aspect	GPT-4o	Gemini Advanced	Claude 3.5 Sonnet
Feedback delivery	Cut-off in mid-output Sometimes, dividing output required.	Cut-off or incomplete feedback. Multiple trials required for a complete output	Output in a single response
Stability across trials	Highly variable owing to GPT's creativity	Some variations, and hallucinations in output	Stable and consistent

The initial comparative training results established Claude 3.5. Sonnet as a preferable choice over other AI platforms. However, beyond these general training outcomes, the analysis of the feedback provided to student papers also supported this final decision by providing important insights into its error identification performance and the readability of content feedback for B1-level students.

In the analysis, the feedback performance of the models was evaluated based on the four essential features. First, accuracy was assessed based on the correct use of error codes

and the relevance of the content feedback provided. This was checked first by researcher, but later to ensure greater accuracy and pedagogical validity, the AI prompt was subsequently reviewed by an expert panel. A second important issue was stability, meaning that the model's output remained consistent across different student papers without significant variation in its delivery format. This was checked with different trials.

Readability was also a crucial factor, with the expectation that feedback would be accessible to B1-level learners; this was verified using tools such as Text Inspector (Weblingua Ltd., n.d.) and the Oxford Readability Index (Oxford University Press, n.d.) In terms of content relevance, the feedback was evaluated on how well it targeted the specific needs of each paper while staying within the word limit, offering a balance of strengths and areas for improvement. It was particularly important that the feedback prioritized the top three content issues and addressed them in a paper sensitive and flexible manner. Finally, the tone of the feedback, especially in error description, was examined to ensure it was constructive and guiding rather than merely corrective.

To summarize the findings, Claude 3.5 Sonnet outperformed the others in error identification, except in cases requiring the RW (rewrite) code because as the other platforms, it performed poor in addressing meaning-level issues. A detailed comparison of the AI models' feedback performances is presented in comparative Table 3.3.

Table 3.3

A Summary of the Comparative Analysis of Feedback Performances Across AI Models

Feedback Quality	GPT-4o	Gemini Advanced	Claude 3.5 Sonnet
Metalinguistic feedback (Guided discovery)	Mixed (often direct corrections)	Often immediate corrections	Consistent guiding tone
Error Description	Some immediate corrections in error description, not allowing for guided discovery	Some immediate corrections in error description rather than guided discovery	Guiding tone in error descriptions
Content focus	-Addressed too many points - some direct answer provision in content feedback	-Addressed too many points -some direct answer provision in content feedback	-Three prioritized points from its angle -Guiding tone in content feedback
Word limit adherence in content feedback	Exceeded (~300 words)	Exceeded (~300 words)	Consistently within limit of (~200 words)

As for content feedback, Claude 3.5 Sonnet remained consistent with the requested word count whereas the other two platforms exceeded the word limit, covering more points than it was supposed to provide. As for the readability of feedback, readability index results were obtained the Oxford Text Checker and Text Inspector showed Claude's feedback was appropriate for B1-level learners, while the readability indexes of GPT-4o and Gemini showed more accessibility for more advanced B2-level students owing to high frequency of B2 words and presence of some C1 words.

Consequently, Claude 3.5 Sonnet was selected as the most suitable AI platform for the context of this study (See Appendix-2 for the detailed training results of three platforms and section 2.4. for researchers' manual analysis of training papers). The steps of the training procedure, including information about the intended feedback, training principles, learning approaches used, and specific steps, are presented as a visual summary in Figure 3.2.

Table 3.4 presents a comparative summary of the feedback readability analysis across the three AI models.

Table 3.4

A Summary of the Comparative Readability Analysis of AI Models

CEFR Level Distribution	GPT-4o	Gemini Advanced	Claude 3.5 Sonnet
Oxford Text Checker Results for Readability	Considerable presence of B1-B2 vocabulary, as well as C1 level (one item)	Moderate presence of vocabulary from both B1 and B2, C1 present	Balanced with A1-A2, as well as presence of B1 vocabulary. No presence of C1 vocabulary
Flesch Reading Ease	47.62/47.13 (B2-C1)	41.89/45.74 (B2-C1)	60.78/59.66 (B1)
Readability score obtained from Text Inspector	No	No	Yes

After weighing all the evidence obtained from the training results, Claude 3.5 Sonnet emerged as the most plausible option among the others. It consistently delivered the desired feedback without altering the output format, provided feedback in a more appropriate tone, and delivered content feedback at an accessible level aligned with the CEFR for B1-level students. However, its output was not flawless because it incorrectly applied rewrite codes or suggested unnecessary rewrites even for some correct sentences,

and its content feedback was template like, sometimes repeating similar points across different responses. These results suggested that, before starting the actual study, the AI prompt needed additional guidance regarding rewrite (RW) cases to help Claude 3.5 Sonnet better detect sentences requiring complete rewrites. Additionally, for content feedback, Claude tended to include similar points that remained consistent across different responses. To address these issues of consistency and flexibility in both error identification and content feedback, two additional measures were incorporated into the training as a part of reinforcement strategy:

- **Using a Pre-Reminder Checklist:** A checklist was developed to address common issues overlooked by the AI during the pilot phase despite the established prompt. This helped reinforce significant points and establish more consistency in results.
- **Shortened Version of the Prompt Uploaded Before Each Paper:** When the content feedback appeared repetitive in points in all papers, a shorter prompt was created to ensure the AI addressed critical feedback components such as the three desired outputs and a guiding, constructive tone. The short prompt included more stress on guiding tone and flexibility in content feedback by prioritizing important needs.

Please provide feedback on the following student essay using the error coding system, error table format, and content feedback guidelines we've discussed. Remember to use a positive and guiding tone in the error descriptions and content feedback, helping the student identify issues without giving away exact corrections. In content feedback, when referring areas of improvement, prioritize the most important three needs for the student's essay.

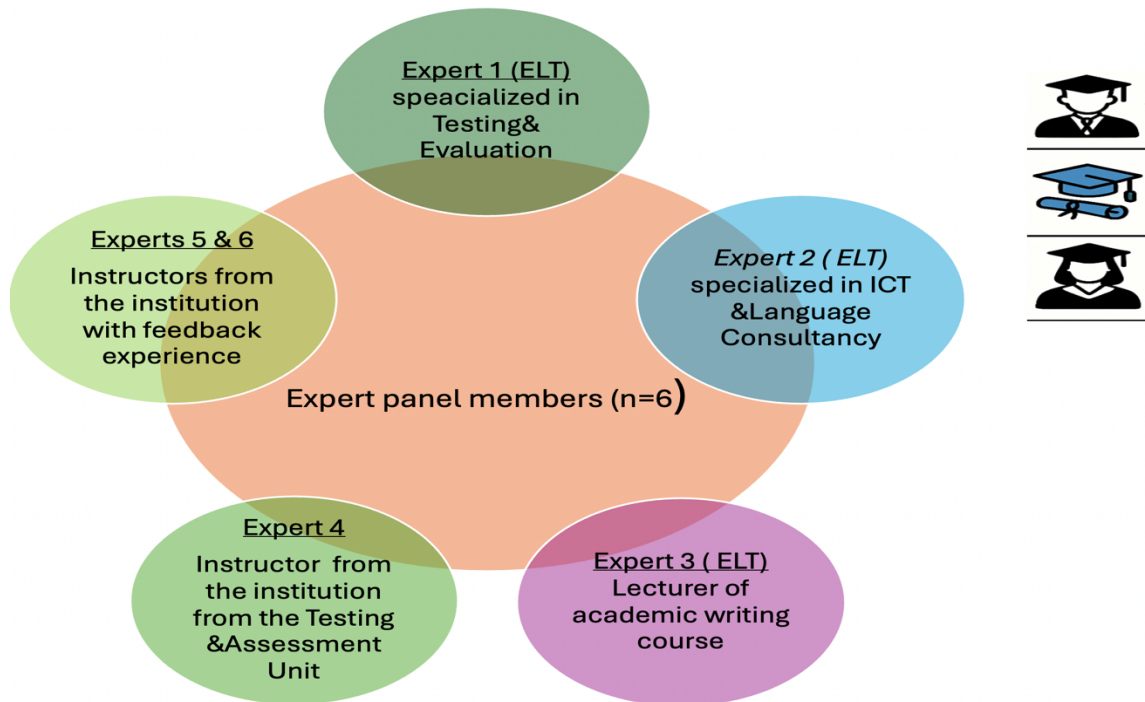
After these revisions and additions, the full implementation started out. (See Appendix-3 for some sample papers used in the actual study to generate AI feedback.)

3.4.3 Expert Review of the AI Prompt and Data Collection Instruments

After training AI platforms, the instruments underwent expert review before starting to train teachers. A panel of six experts reviewed both the AI prompt and data collection instruments via a Google Form. Based on their suggestions, a finalized version was created (See Appendix-4 for the output of received expert opinions). The expert panel consisted of professionals with diverse academic and practical expertise as presented below in Figure 3.3.

Figure 3.3

Composition of the Expert Review Panel



- One associate professor specializing in Testing and Evaluation
- Two PhD holders with academic backgrounds in English Language Teaching (ELT); one is an expert in Language Consultancy and Information and Communication Technology (ICT), while the other is currently working as an instructor in an ELT department.
- Three instructors holding master's degrees, one of whom is a member of the Assessment and Evaluation Department at a School of Foreign Languages, and the other two have practical experience with the institution's feedback system.

Overall, these different areas of expertise helped ensure that the tools and procedures used in the study were carefully reviewed from both teaching and technology perspectives.

3.4.4 Training of Teachers

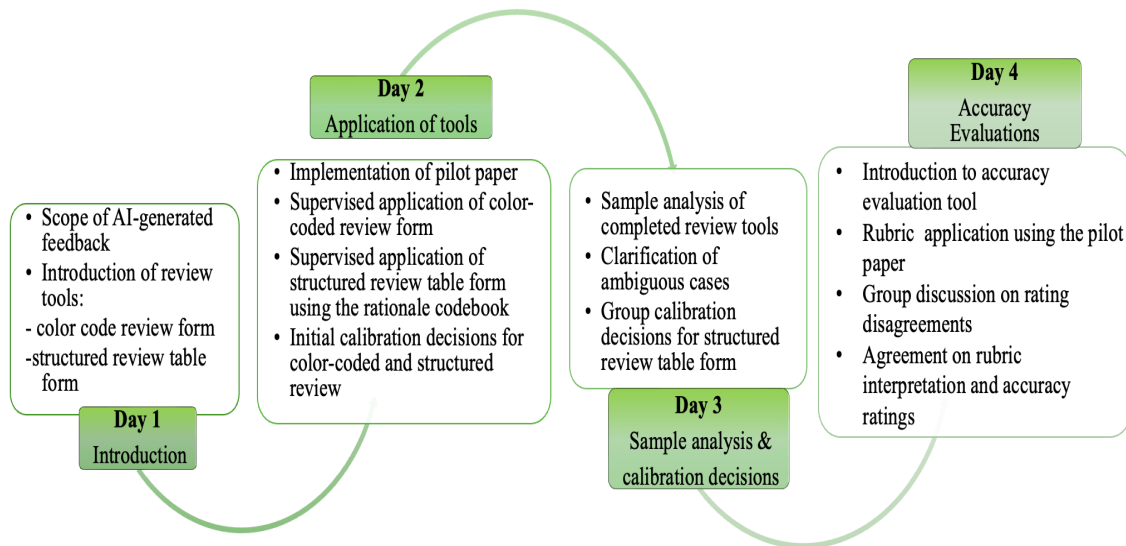
Once the expert-reviewed instruments were finalized, teachers underwent training to ensure consistency in AI feedback evaluation. The training materials, including powerpoint presentation, materials and a teacher refinement sample used in pilot study are available in the Appendix-5. The training familiarized teachers with AI-generated feedback output and provided hands-on experience in:

- Reviewing and refining AI feedback with color-coded form.
- Documenting acceptance levels, changes, and rationales in the structured review table form.
- Evaluating AI-generated feedback using the accuracy evaluation rubric developed for this purpose.

The following schedule presented in Figure 3.4 provides a brief overview of training and content of the program:

Figure 3.4

Overview of Training Schedule



Day 1: Introduction

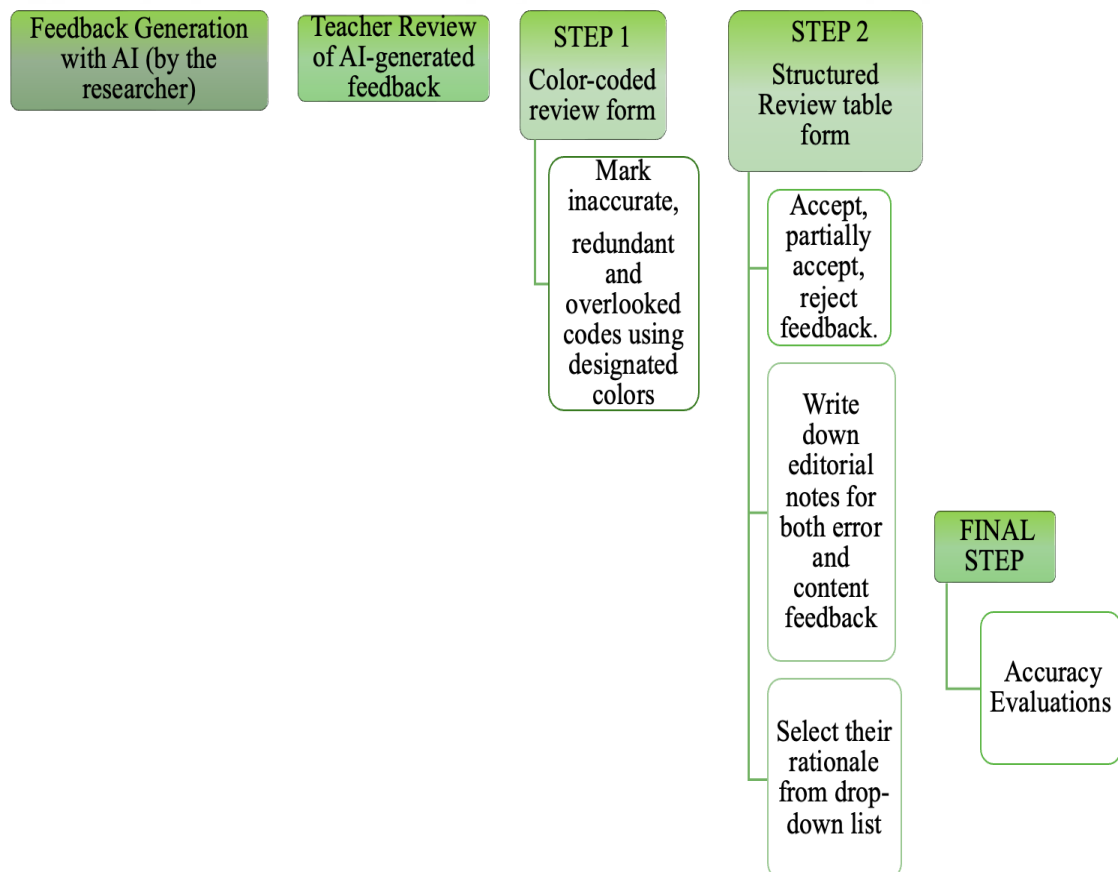
In first day of the training, teachers were provided with a quick overview of the research aim and an introduction to AI-generated feedback. During this session, teachers were also informed about the scope of AI feedback and were asked to examine feedback points in the sample AI feedback to set expectations for AI-generated feedback. After that, they were given a quick introduction to the research tools for reviewing and evaluating AI feedback.

Besides, teachers were provided with a general guide for conducting color-coded review and completing structured review table, to help them understand the basics of reviewing and refining AI feedback with these review tools. The color-coded review was explained to participants with a demonstration of a sample analysis. The structured review

document was also demonstrated, and the same general guideline and the code book for selection of rationales were provided to teachers as a handout. At this stage, teachers were explained the meanings of each rationale and reached common understanding as to how to use these codes. Additionally, to emphasize critical importance of structured review document, teachers were informed that the acceptance levels in the structured review table can also serve as the foundation for consistent and reliable rubric ratings across raters. During this session, the following review workflow process was shared with teachers for them to follow them systematically, as presented in Figure 3.5.

Figure 3.5

Workflow for Teacher Refinement process and task completion



Day 2: Reviewing and refining Sample AI feedback as a part of pilot study

On the second day of training, teachers were trained with a pilot paper to gain hands-on experience and to pilot the instruments before full scale implementation of them. In the color-coding review phase, teachers were provided with error-incorporated student papers to refine, and they were asked to practice applying the predefined color-coding system to student essays and logging their acceptance levels of AI feedback, rationale codes and

their notes for implemented changes in the corresponding structured review table used for documenting these.

The review assignment was carried out by the teachers during the session under the researcher's supervision. This session served as a valuable opportunity to establish a standard procedure for the review process together. During the activity, teachers suggested several important additions and clarifications to improve the practicality of the tools used. As a result, some calibration decisions were taken for both the color-coded reviews and the structured review tables which were to be employed during the actual study.

Day 3: Analysis and discussion of Sample Review Tables Completed by Teachers and Peer Review (Brief meeting to clarify doubts)

The third day session focused on the teachers' sample review tables, which were analyzed by the researcher to clarify unclear aspects of the review process. Based on this analysis, teachers were reminded of the distinctions between each color-coding category and were guided on how to select appropriate acceptance levels based on their chosen rationale. The researcher also debriefed the teachers that, to ensure their pedagogical reasoning could be accurately interpreted, they needed to refine their feedback in a careful and thoughtful manner by documenting all changes and explaining suggested modifications with clarity. They were also informed that their reviews would be monitored and cross-checked, and hence, if they experienced any confusion during the review process, they were asked to contact the researcher in person or via e-mail. Following this session, teachers were provided with the assessment rubric for accuracy evaluation, which they were instructed to read before attending the final session.

Day 4: Accuracy evaluations

The final day of training focused on the employment of assessment rubric of accuracy which was designed to evaluate the AI-generated feedback. Teachers were introduced to the evaluation criteria and scoring procedures, and it was explained that their acceptance level selections in the structured review documentation could help guide their overall holistic judgments. During the session, teachers evaluated a sample paper and then shared their rationales. Some participants expressed uncertainty about whether they had assigned the correct ratings, so they were asked to cross-check their selected rationale codes, which helped them better judge the severity of the errors.

Following this session, it was decided to include quick-reference default and mixed scenarios corresponding to each acceptance level. This aimed to enhance the practicality and consistency of scoring across raters.

As a final step, teachers completed a training evaluation form, in which they shared their opinions on the training process and the instruments used, and recorded their final suggestions before the actual study began.

3.4.5 Refinements implemented after the pilot study

As discussed above, the pilot study was integrated into the teacher training process to test and refine the review tools prior to the full implementation. The feedback from eight instructor participants was gathered via Google Form, and their suggestions led to improvements in the following instruments:

- **Color-coded review form:** a Word file where teachers applied the designated color-coding system to evaluate AI-generated feedback. For example, red was used for incorrect codes, and blue for overlooked errors.
- **Structured review form:** an excel sheet including review tables where teachers record their acceptance levels (Accept, Partially Accept, Reject), documented the changes they would make to AI feedback, and choose selected rationale codes to justify their decisions.
- **Accuracy evaluations with (rubric):** a three-scale holistic evaluation rubric to evaluate AI feedback quality based on acceptance levels recorded in the structured review table.
- **Semi-structured interview questions:** questions aim to gauge teachers' perceptions of accuracy of AI feedback and experiences with AI feedback during refinement process, as well as their perceptions for future integration of AI feedback in classrooms.

The adjustments which were implemented to the data collection tools were summarized in the headings below:

3.4.5.1 Adjustments for the color-coding review form

Color-coding feedback form were adjusted based on teachers' needs emerge after their review process during pilot study. For instance, teachers recommended introducing an 'RW' (rewrite) code to distinguish between partial and full rewrites:

- Partial rewrites: Underlined within the sentence.
- Full rewrites: Marked with RW at the end of the sentence.

Additionally, teachers suggested removing unnecessary color coding for content feedback and instead documenting changes directly in the structured review table to streamline the review process and avoid double-work.

Updated Color-Coding Scheme:

- **Red** → Incorrect AI feedback (wrong error code placements).
- **Blue** → Errors overlooked by AI.
- **Yellow + X** → Redundant AI-generated codes.
- **RW (underlined text)** → Partial rewriting required.
- **RW (at end of sentence)** → Full sentence rewriting needed.

3.4.5.2 Adjustments for structured review table form

Just like color-coding scheme, Teachers also suggested additions for the severity of error, adding a category for major overlooked errors (e.g., missing main verbs, auxiliary verbs, rewriting needs). Based on their feedback, the following rationale numbering system was established for the review process:

Table 1: Error-Code Feedback

- **Low Accuracy (Reject 1)** → Rationale numbers 1–6.
- **Mid Accuracy (Partially Accept 2)** → Rationale numbers 7–14.
- **High Accuracy (Accept 3)** → Only one item, numbered 15.

Table 2: Content Feedback

- **Low Accuracy (Reject 1)** → Rationale numbers 16–18.
- **Mid Accuracy (Partially Accept 2)** → Rationale numbers 19–22.

- **High Accuracy (Accept 3)** → Only one item, numbered **23**.

Teachers also suggested two things to improve practicality and consistency of review process:

- **Simultaneous review using both forms:** Instead of separate review phases, teachers preferred to perform the color-coding review and structured review table simultaneously to improve accuracy and time-efficiency.

Quick Numbering Combinations for Rationale Selection: To prevent missing rationale numbers, teachers identified co-occurring categories: No. 1 with No. 4; No. 8 with No. 11; No. 7 with No. 11. Teachers were explained that selecting one of these often implies the need for the other.

3.4.5.3 Adjustments for the Accuracy Evaluation (rubric)

During the pilot training, teacher also tested the three-point scale accuracy evaluation rubric for robust evaluation, and accordingly, they suggested two key improvements for both practicality efficiency and maintain consistent rating across raters.

- **Default Scenarios:** The addition of default and mixed scenarios of ratings to facilitate decision-making during evaluations.
- **Accuracy evaluations in one form for all:** Instead of working on separate documents, rubric evaluations were conducted within a shared OneDrive file containing all ten papers.

3.5 Data Collection Tools

Data collection tools, including the AI prompt, were validated by experts and subsequently tested in a pilot study to further ensure the effectiveness of the instruments before full-scale implementation. . In total, four tools were used: a color-coded review and a structured review form to collect data on how teachers refined AI feedback; a rubric to evaluate the accuracy of that feedback; and semi-structured interviews to gather teachers' insights on its accuracy, classroom usability, and suggestions for improvement.

3.5.1 Color-coded review form

The color-coded review was developed to visually display teachers' refinements on AI-generated feedback by marking incorrect, redundant, or overlooked feedback directly on

the student essay. This approach made teacher-made corrections easily traceable and transparent (See Appendix-6 for the color-coded review guidelines).

3.5.2 Structured review table form

The structured review table was designed to help teachers systematically document their feedback refinements and enhancements on AI feedback in a user-friendly and systematic way. This allowed researchers to do close monitoring and cross-check of data and intervene possible problems (See Appendix-7 for structured review table). Each table included:

- AI raw feedback inserted in the form by the researcher
- Drop-down menus for teachers to select their acceptance level (Accept, Partially Accept, Reject)
- A section for Teacher notes
- Drop-down lists with rationale codes, and the “other” section for other types of emerging rationales

These rationale codes were developed based on recurring issues identified during the prompt training of Claude 3.5. Sonnet, GPT-4o, and Gemini Advanced. The patterns observed in the AI-generated outputs informed the design of the review table, ensuring that it would support both teacher’s documentation of feedback issues.

3.5.3 Accuracy Evaluations

A structured three-point rating scale of evaluation tool was developed to holistically evaluate the accuracy and quality of AI-generated feedback, as well as to gauge the level of teacher intervention needed (see Appendix-8 for the accuracy evaluation guidelines). The design of the scale was informed by the principles outlined by Popham (2007), which emphasize establishing clear evaluation criteria, defining those criteria, and providing an actionable scoring strategy.

In addition to experts’ review, teacher input during the pilot phase played a significant role in the refinement of the rubric into its last form. Based on their evaluations, adjustments were made, and this supported clear scoring of the rubric for teachers. The three-point scoring scheme is presented below:

Inaccurate Feedback/ Low accuracy (1): Feedback contains incorrect error identification, major overlooked errors, misleading suggestions, or completely wrong and unhelpful explanations. (e.g., assigning the wrong error code to a sentence or failing to provide constructive (overtly directive tone) and guidance for self- correction (providing direct answer/ solution for the error)).

Partially Accurate Feedback/Mid accuracy (2): Feedback correctly identifies errors but provides unclear, incomplete with minor overlooked errors, or partially misleading or vague error explanations. (e.g., marking a word error correctly but failing to specify if it is word form or word choice).

Accurate Feedback/ High accuracy (3): Feedback correctly identifies the error and provides a clear, constructive explanation with appropriate guidance. (e.g., accurately identifying a subject-verb agreement mistake and explaining the rule with a relevant self-correction suggestion (i.e. consider if the plural subject match with the verb)).

To enhance the reliability and practicality of teachers' evaluations, the rubric also includes two other additional components. These are intended to support teachers' decision-making by enabling them to conduct a more consistent and confident evaluation, especially in cases where they may feel uncertain about their final judgment. These two components were integrated to promote more valid and reliable rating by constituting written and visual anchor examples for teachers:

- **Pre-rubric guidance**, helping teachers make consistent and valid judgments by allowing teachers to reference examples and clarify uncertainties
- **Default and mixed scenario ratings**, which teachers can use to do evaluation more practically.

Literature suggests the use of anchors, such as annotated examples or actual work samples help illustrate reviewers' different levels of performance better (Brookhart, 2013). In the present study, following this practice, pre-rubric guidance, scenarios, and a sample evaluation were used to help teachers understand the intricacies of evaluation and ease their decision-making.

3.5.4 Semi-structured Interview

Semi-structured interview questions were also formed to delve into teachers' perceptions of AI-generated feedback based on their filtering and refining experiences (See

Appendix-9 for interview Questions). For this reason, the interview was structured in three parts: (1) teachers' perceptions of AI feedback accuracy, (2) pros and cons of the refinement process, and (3) views on the future use of AI in the classroom along with their suggestions.

The aim of the interview was to encourage teachers to elaborate on the key aspects they found particularly significant, while also ensuring consistency across participants' responses. To achieve this, special attention was paid to the clarity, neutrality and focus of the questions (Turner, 2010). Accordingly, expert opinions were gathered from the same group of specialists via Google Form. Based on their recommendations, the questions were revised for greater clarity and neutrality in question wording, eliminating unnecessary complexity and making them more targeted and concise. This was intended to help teachers recall specific refinements and experiences from their review processes and effectively communicate their insights. To further support memory recall, , teachers were encouraged to revisit their reviews before the interview and keep them on hand during the session as a reference. Each interview was conducted individually and lasted approximately 10 minutes.

The summary of all data collection tools with their descriptions, and their contribution to research questions are explained in below in the Table 3.5.

Table 3.5

Summary of Data Collection Tools, Descriptions of tools, and Contribution to Research Questions

Data Collection Tools	Description of Tools	Related Research Question(s)	Contribution to Research Questions	Data Analysis Approaches
Color-coded review form	Teachers revised AI feedback directly on student papers using color codes (e.g., red = incorrect codes, blue = missed codes, yellow = unnecessary codes).	RQ1	Contributed to RQ1 by identifying teacher-led refinements through color-coded corrections.	Summative + Conventional Content Analysis: Counted color-coded categories; interpreted teacher notes.
Structured review table form	Teachers selected acceptance levels (Accept, Partial, Reject), wrote editorial comments, and chose rationale codes. They also refined content feedback within this form.	RQ1, RQ2	Contributed to RQ1 (how teachers refined AI feedback) and RQ2 (perceived accuracy levels via acceptance decisions).	Summative + Conventional Content Analysis: Counted acceptance decisions and rationale codes; analyzed teacher comments.

Table 3.5(continued)*Summary of Data Collection Tools, Descriptions of tools, and Contribution to Research Questions*

Data Collection Tools	Description of Tools	Related Research Question(s)	Contribution to Research Questions	Data Analysis Approaches
Accuracy evaluations (rubric)	Teachers rated AI feedback using a 3-point accuracy rubric and added qualitative comments.	RQ2	Contributed to RQ2 by providing holistic accuracy ratings and explanatory insights into teachers' judgments.	Consensus rating calculation + Qualitative interpretation of rubric comments.
Semi-structured interviews	One-on-one interviews with open-ended questions about experiences with AI feedback.	RQ3	Contributed to RQ3 by capturing reflections on benefits, challenges, and classroom applicability.	Conventional Content Analysis: Emergent themes coded and grouped.

As shown above, each data collection tool serves differently to the purpose of the present study and complement each other's finding. And in the analysis of the results, including the combination of both summative counts and conventional content analysis enabled to reach a comprehensive understanding of how teachers refined and evaluated AI-generated feedback.

3.6 Data Analysis

The present study employs a qualitative research approach using a combination of summative and conventional content analysis methods as defined by Hsieh and Shannon (2005). According to Hsieh and Shannon (2005), there are three distinct approaches to content analysis: conventional, directed, and summative. While conventional analysis works with raw data when literature is limited, directed analysis is guided by existing literature. The summative approach combines quantitative and qualitative method, where the two-fold analysis begins with quantifying specific words or content (manifest content) to establish a baseline, then moves to qualitative interpretation of the underlying themes (latent content).

In this respect, the combination of summative and conventional content analysis methods, as defined by Hsieh and Shannon (2005), was selected based on the nature of the data in

the present study. Both methods were applied to different data types to capture both common patterns and themes.

In order to review AI-generated feedback, teachers refined AI feedback-generated feedback through two structured review forms: (1) a color-coded review of AI-generated error feedback and (2) a structured review table documenting; acceptance levels, rationale codes, brief editorial comments for error-coded and content feedback. Summative content analysis was applied to color-coding and structured review forms to identify types of feedback refinements, acceptance levels, and patterns in rationale use. Teachers' short notes for edits were not included in the count, but they were qualitatively examined to perform meaningful interpretation of teachers' reviews.

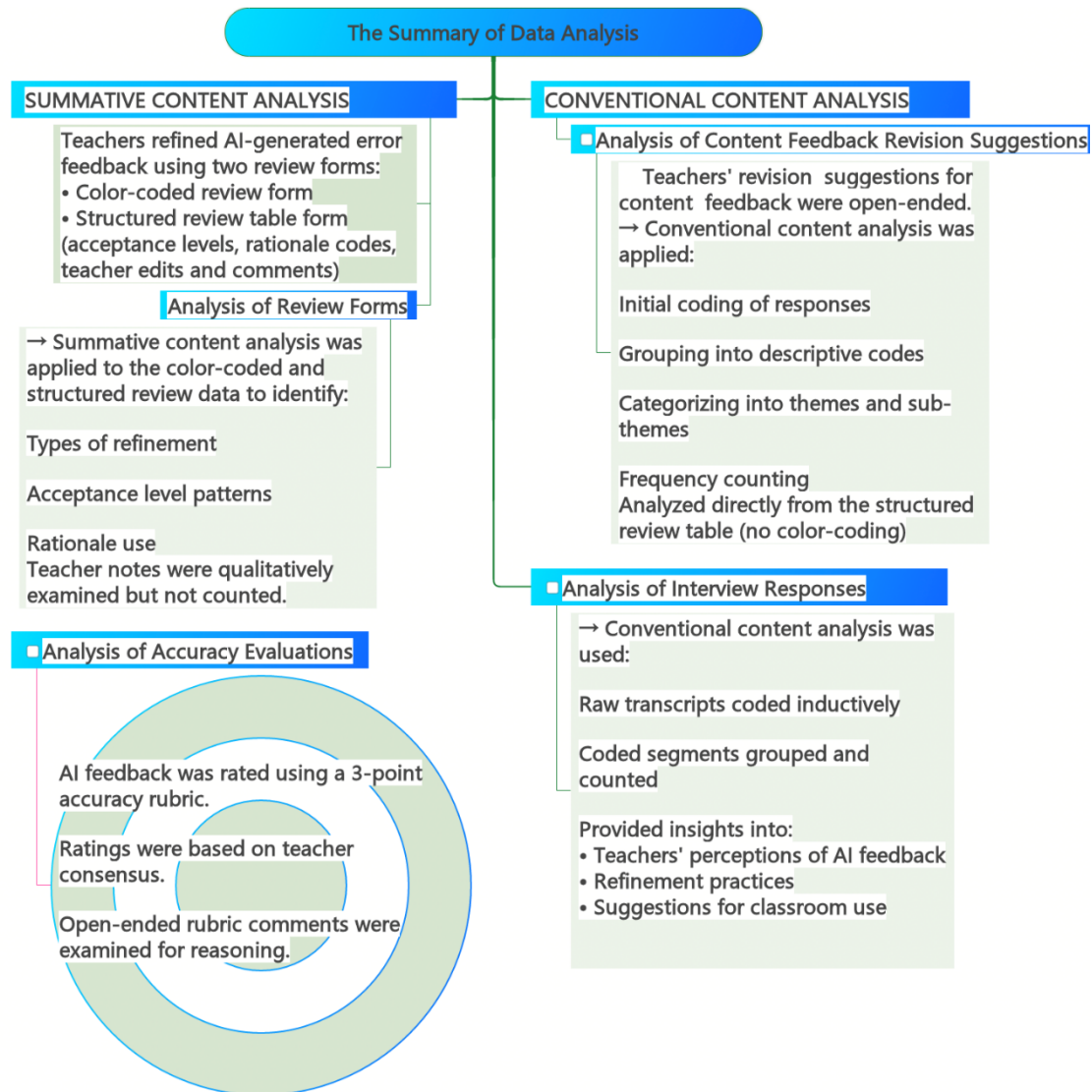
For content feedback, conventional content analysis was used, as the teachers' commentary and suggestions were more open-ended. These were reviewed and assigned preliminary codes, which were then grouped into descriptive codes, categorized into emerging themes and sub-themes, and counted to take frequency. Unlike error feedback, this part was not color-coded, but they were analyzed directly from the structured review table. Similarly, interview data were analyzed using conventional content analysis, as the codes emerged inductively from the raw data.

Accuracy evaluations provided a separate layer of analysis. Each teacher rated the accuracy of AI-generated feedback using a 3-point rubric for both error-coded and content feedback. Accuracy scores were determined based on consensus ratings agreed upon by teachers. Additionally, open-ended rubric comments were also examined to contextualize teachers' reasoning and concerns in greater depth.

Interview data were also analyzed through conventional content analysis where data elicited inductively from the raw transcripts only. For interview analysis, segments were highlighted, coded with preliminary codes, grouped into descriptive categories, and counted to gain insight into teachers' perceptions of AI feedback, their refinement practices, and their suggestions for classroom use. Figure 3.6 below illustrates how summative and conventional content analysis methods were applied to different types of data, including error-coded feedback, content feedback, accuracy evaluations, and interview responses.

Figure 3.6.

The Summary of the Application of Summative and Conventional Content Analysis Methods to the Data



3.6.1 Overview of the data collection procedure

The data collection procedure was planned to be conducted over a 12-week period. Accordingly, data were gathered in the Fall semester (2025) of the 2024–2025 academic year. Prior to data collection, the AI prompt and comparative training results were finalized.

Before starting actual study, teacher training and piloting were conducted to prepare teachers for full scale implementation. Teachers began their reviews of AI-generated feedback towards the end of the second week. During their reviews and evaluations, multiple calibration and standardization sessions were held to ensure consistency in teachers' reviews and evaluations. Finally, individual interviews were conducted in the

last stage of the process, immediately after the reviews were finalized. The overview of data collection procedure is presented in Table 3.6 below.

Table 3.6

Overview of the Data Collection Procedure

Weeks	Activities	Descriptions
Week 1	AI Feedback Generation by the researcher	AI-generated feedback was created using Claude 3.5 Sonnet.
Week 1-2	Training & Initial Standardization	Training session and initial standardization meeting conducted (pilot study with 1 paper) to familiarize teachers with the review process and rubric evaluation.
Week 3	Calibration session 2	Calibration session for clarifying doubts and remaining consistent in analyses.
Week 4	Paper Review 4	Teachers reviewed paper 4
Week 5	Brief Standardization & Mid-point Calibration 3	Mid-point standardization meeting and calibration session to realign evaluators if needed
Week 6-8	Paper Review (6-8)	Teachers reviewed papers 6 through 8
Week 8	Brief meeting for any lingering questions	Brief meeting held to address any remaining teacher questions or clarifications.
Week 9-10	Completion of their reviews	Teachers completed all of their reviews.
Week 11-12	Interviews	Teachers conducted individual interviews with teachers

3.6.2 Researcher reflexivity

As a teacher my competence as a teacher-user of AI was limited to lesson plans and generating activities for my classes, all of lessons, including writing. However, in time, as I became more confident and apt in using AI, I gradually expanded to providing small feedback notes on student writing. At that moment, it hit me the realization that maybe

AI could be used for feedback on student essays, a task that often overwhelms me and my colleagues due to the number of students and time constraints.

Besides challenge of providing feedback in time constraints to large number of students, student profile was also another factor challenging our provision of feedback. The significant amount of our students are Turkish learners, which leads them to write essays with frequent grammar errors, language transfer issues, and structural differences in their writing. Many students are in tendency to translate directly from Turkish, thinking in their native language, which causes unnatural phrasing and syntactical inconsistencies while writing essays in English. As a result, their papers require feedback that not only corrects errors but also guides them in expressing ideas more naturally and contextually appropriately in English. Managing these multi-layered feedback needs within a limited time frame creates additional stress for teachers.

This pedagogical challenge urged me to explore whether AI could support feedback practices by tailoring its language and tone to be teacher-like and incorporating the coding scheme and metalinguistic elements we have been using in our classrooms for B1 level writing students. With that in mind, my research quest and inquiry ventured forward. I was hoping that even if AI failed, I could still gain valuable insights from teachers that would unfold what AI is capable of and what it is not.

Although I was experienced in providing written feedback to students, I had minimal knowledge of how to effectively interact with AI tools to generate feedback aligned with pedagogical expectations. This gap became a driving force in my learning during the AI training process. I began investigating various tools to identify the one most capable of meeting my research goals and generating the desired type of feedback, which led me to delve deeper into the world of AI. This started with a robust literature review on the capabilities and limitations of GenAI tools and comparative studies on the tools I adopted for the present study. This gave me insight into design natures, general capabilities, and feedback generation features.

Through my trials and frustrations, I came to understand both the reasoning behind AI responses and thus discovered the constraints inherent in their design. During the development of prompts, I tried different prompt types, compared outputs of three tools, and reported their findings. I wrote reflections on what worked and what did not, and as a result of that, created a pre-reminder checklist for AI to respond better to checklist items

where it had previously shown oversights. Thus, my training started as a novice user but evolved into an informed and critical user.

The participating teachers in this study were my colleagues from the same institution. Their role was to review, refine, and assess the AI-generated feedback rather than directly using AI themselves. At the onset, they felt excited about this venture and volunteered to help. Like me, they had limited experience using AI and did not fully know its capabilities regarding writing feedback. However, they had reservations due to past inaccuracies they had encountered while using AI in classroom activities. They were also aware of the need for double-checking and refining AI-generated feedback for greater accuracy.

Throughout this research, I remained conscious of my dual role as an observer and researcher. My focus was on analyzing how teachers interacted with AI-generated feedback and identifying patterns in their refinements. Engaging in this research as a researcher helped me remain objective, ensuring that my interpretations were guided only by teachers' evaluations and empirical findings rather than my own personal assumptions.

3.7 Trustworthiness of the Study

The present study employed qualitative research criteria for ensuring trustworthiness, including credibility, transferability, confirmability, and dependability, as outlined by Lincoln and Guba (1985).

Creswell (2016) notes that triangulation of data serves as an essential initial step in establishing the credibility of a qualitative study. In this vein, data of this study were gathered from four different main sources complementing and ensuring credibility: (1) color-coded review, (2) structured review table, (3) rubric-based evaluations of AI-generated feedback, and (4) follow-up teacher interviews. This variety of sources enables the development of credible interpretations and cross-checking of findings. In addition to methodological triangulation, an expert review team composed of experienced English language teaching professionals evaluated all data collection tools and processes, which also contributed to the credibility of the data gathered from the developed tools.

Aside from triangulation and expert review, multiple other strategies were employed for the trustworthiness of the results. For the refinement reviews, a pilot paper was used at the start to familiarize teachers with the refinement review process. Additionally, in order

to standardize the review procedures, calibration sessions were conducted at the beginning, midpoint, and follow-up check-ins were made when needed.

As for the credibility of rubric evaluations, all teachers evaluated the pilot paper using the same rubric to establish a shared understanding of how to apply acceptance levels (Accept, Partially Accept, Reject). In addition, three meetings were held with teachers during data collection to discuss disagreements and clarify rubric interpretations, which enabled inter-rater consistency. In the interpretation of results, consensus ratings were used to ensure interpretations were free from researcher bias. If most teachers rated an item as "mid-accuracy," this became the consensus rating.

For the interview data, member checking was implemented to ensure the credibility of the findings. In this vein, interview transcripts and coding table were sent to participants for verification, and necessary revisions were made to reflect their intended meanings (Creswell & Poth, 2017). Additionally, a fellow colleague also reviewed the codes and emergent themes to ensure intercoder reliability and strengthen trustworthiness (Nowell et al., 2017).

Transferability of the results was supported through a thick description of the research context, including the institutional setting (a university preparatory program), participant profiles (experienced English language instructors), and the types of AI-generated feedback evaluated (error-coded and content-level feedback on student essays). The procedures for reviews and evaluation were described in detail to enable other researchers to assess the relevance of these findings to similar educational settings.

As for confirmability, this study ensured clear documentation of the analysis process to ensure data obtained from teachers' input and not influenced by researcher bias. For the teacher review and rubric evaluation phases, an audit trail was maintained, including records of calibration discussions and review decisions, which were shared with the teachers. This transparency allowed teachers to follow all decisions taken during process and teachers to do their review in an uniformed way. The researcher contacted participants to clarify their intended meaning when any review was unclear or incomplete, rather than making assumptions. On the other hand, teachers could also reach out to the researcher to ask follow-up questions as needed, which helped ensure mutual understanding. Similarly, for the rubric calibrations, the same transparency and shared documentation allowed for a clear and traceable process for teachers. As for interviews,

the coding and themes were grounded in teachers' responses, with a traceable coding framework for preliminary and descriptive coding.

Dependability was also maintained through standardized procedures throughout data collection, such as following the same evaluation protocol and uniform calibration decisions. To quantitatively verify the dependability of rubric evaluations, a Coefficient of Variation (CV) analysis was conducted, yielding an overall CV of 7.89%. The CV for low accuracy ratings was 6.12%, for mid accuracy 8.34%, and for high accuracy 0.00%, reflecting perfect agreement. As all values were below 10%, these results confirm the consistent application of the rubric across the dataset.

Finally, to mitigate social desirability bias, participants were assured that their identities would remain anonymous for confidentiality. Besides, interview questions were carefully reviewed for neutrality, so as not to lead them into particular answers. Moreover, the researcher created a non-threatening environment for the honest expression of views (Patton, 2015).

This systematic approach strengthens all four aspects of trustworthiness: credibility through shared interpretations and member checking; transferability through documented processes; dependability through consistent application of criteria and co-rating; and confirmability through minimization of individual biases and cross-checking procedures.

3.8 Ethical Considerations

In accordance with ethical research principles, several measures were taken prior to the implementation of this study. First, ethical approval was obtained from the Ethics Committee of Anadolu University before data collection began (see Appendix-16). In addition, all participants were fully informed about the purpose of the study, their voluntary participation, and their right to withdraw at any time. Informed consent forms were distributed, which clearly explained the research objectives, data collection procedures, and the participants' contribution to the study. All participants signed the form to provide their official consent (see Appendix 17).

To ensure confidentiality, participants were assured that their identities would remain anonymous. In the findings section of this thesis, pseudonyms in the form of codes (e.g., T1, T2, etc.) are used instead of their real names. Furthermore, all collected data were stored in a password-protected digital environment to ensure privacy and data security.

4 FINDINGS

This chapter presents the findings of the study investigating how preparatory class teachers refine and evaluate AI-generated feedback on student essays. The research aimed to understand how teachers refine AI-generated feedback and their evaluations of AI-generated feedback accuracy based on their refinements. Also, based on their experiences of refining, the study also explores teachers' disposition towards refining and its classroom usability, and provide suggestions to improve it. Data were collected through a content analysis of multiple data sets, including teachers' color-coded review forms and structured review tables, in order to explore their refinements and accuracy evaluations of AI-generated feedback. Additionally, semi-structured interviews were conducted with eight university English language instructors to gather deeper insights into their perceptions and experiences. The collected data were then subjected to summative and qualitative content analysis based on the nature of data, as explained in detail in the Methodology section.

4.1 Findings related to Research Question 1

The first research question aimed to analyze how teachers refined AI-generated feedback with their own expertise and insights. This section delves into the ways in which teachers enriched and reworked AI-generated feedback by drawing on their pedagogical expertise and professional judgment. The analysis covers both error-coded and content-related feedback, along with the rationale behind teacher-implemented refinements. Representative samples of teachers color-coded and structured reviews used in the refinement process are provided in Appendix-10. The subsequent subsections offer a detailed account of the quantitative and qualitative findings, beginning with error-coded feedback, followed by an analysis of content feedback.

4.1.1 Quantitative analysis of teachers' refinement patterns for error-coded feedback

In order to analyze teachers' refinements systematically, Teachers refined AI-generated error-coded feedback by using color-coded scheme, and address mainly three main types of issues: inaccurate code, redundant code, and overlooked code, which enabled understanding common types of errors emerged from teachers' reviews. The refinement color-coded for the review process were presented below:

- **Inaccurate Codes** (marked in red): Instances where AI incorrectly identified the error type
- **Redundant Codes** (marked in yellow): Instances where AI provided redundant feedback.
- **Overlooked Codes** (marked in blue): Errors that AI failed to detect but were identified by teachers

The analysis of teachers' refinements across nine student papers revealed a clear pattern in the distribution of refinement types, as shown in Table 4.1.

Table 4.1

Distribution of Teachers' Refinements for Error-Code Feedback Across All Papers

Refinement Type	Total Count	Percentage	Key Pattern
Inaccurate Codes 	180	11.3%	Teachers corrected misidentified error types
Redundant Codes 	171	10.7%	Teachers removed unnecessary or false feedback
Overlooked Codes 	1,244	78.0%	Teachers added missing errors
Total Refinements	1,595	100%	Teachers primarily filled in gaps where the AI gave no feedback

As demonstrated in Table 4.1, the total number of teacher-made refinements was 1,595. The overwhelming majority of refinements (78.0%) addressed overlooked errors that the AI had failed to identify. In contrast, inaccurate codes (11.3%) and redundant codes (10.7%) occurred at similar, much lower rates across the papers.

These findings indicated that the primary limitation of AI feedback was related to its failure to detect a large number of errors across papers. This also points to a fundamental lack of comprehensiveness in the AI's processing, particularly when confronted with multiple or complex issues altering meaning in student writing.

While all teachers focused heavily on correcting overlooked issues, their individual approaches to refining existing feedback still showed some variation, as shown in Table 4.2.

Table 4.2*Refinement Distribution across Teachers*

Teacher	🔴 Inaccurate	🟡 Redundant	🟢 Overlooked	Total	% Overlooked
T1	23 (11.6%)	14 (7.1%)	161 (81.3%)	198	81.3%
T2	20 (9.8%)	24 (11.7%)	161 (78.5%)	205	78.5%
T3	16 (10.3%)	24 (15.5%)	115 (74.2%)	155	74.2%
T4	24 (11.0%)	23 (10.6%)	171 (78.4%)	218	78.4%
T5	21 (11.6%)	23 (12.7%)	137 (75.7%)	181	75.7%
T6	20 (9.0%)	19 (8.5%)	184 (82.5%)	223	82.5%
T7	31 (14.3%)	22 (10.1%)	164 (75.6%)	217	75.6%
T8	25 (12.6%)	22 (11.1%)	151 (76.3%)	198	76.3%

The data reveal that, although all teachers concentrated the majority of their refinements on overlooked errors (ranging from 74.2% to 82.5% of their total modifications), notable differences emerged in how they handled the existing AI feedback. For example, T7 placed greater emphasis on correcting inaccuracies, with 14.3% of their refinements addressing misidentified error types. In contrast, T3 focused more on eliminating redundant feedback, allocating 15.5% of their changes to removing unnecessary or incorrectly applied codes. Meanwhile, T6 recorded the highest number of overall refinements, with an overwhelming 82.5% dedicated to addressing overlooked feedback.

These variations highlight that, while all teachers shared a common objective of compensating for the gaps in AI-generated feedback, their specific approaches reflected differing pedagogical orientations and feedback philosophies.

In line with these results, the distribution of refinements also varied substantially across individual student papers. However, unlike the previous table, this table highlighted how individual teachers differed in their approaches to AI feedback, and that the nature of the student papers themselves also influenced those decisions. The extent of this variation is presented in ascending order in Table 4.3.

Table 4.3*Teachers' Refinement Distribution by Paper*

Paper	 Inaccurate	 Redundant	 Overlooked	Total	% Overlooked
Paper 4	29 (9.5%)	28 (9.2%)	247 (81.3%)	304	81.3%
Paper 5	14 (7.3%)	18 (9.4%)	160 (83.3%)	192	83.3%
Paper 3	20 (10.6%)	23 (12.2%)	145 (77.1%)	188	77.1%
Paper 7	25 (14.1%)	23 (13.0%)	129 (72.9%)	177	72.9%
Paper 10	42 (23.9%)	17 (9.7%)	117 (66.5%)	176	66.5%
Paper 9	19 (11.0%)	17 (9.9%)	136 (79.1%)	172	79.1%
Paper 2	11 (7.5%)	13 (8.8%)	123 (83.7%)	147	83.7%
Paper 8	14 (10.4%)	19 (14.1%)	102 (75.6%)	135	75.6%
Paper 6	6 (5.8%)	13 (12.5%)	85 (81.7%)	104	81.7%

As can be seen from the table 4.3, teacher intervention varies across different student papers based on the linguistic profile of the paper. The table above showed that Paper 4 required the highest number of total refinements (304), whereas Paper 6 involved the fewest (104), suggesting that papers with a higher density or complexity of errors tend to challenge AI systems more significantly and therefore demand greater teacher involvement.

Paper 10 stood out with the highest percentage of inaccurate feedback (23.9%), highlighting that certain complex sentence structures or error types may be more difficult for the AI to interpret accurately. In contrast, Papers 2 and 5 exhibited the highest proportions of overlooked errors (both exceeding 83%), indicating that these texts might have included more subtle or nuanced issues that escaped AI detection.

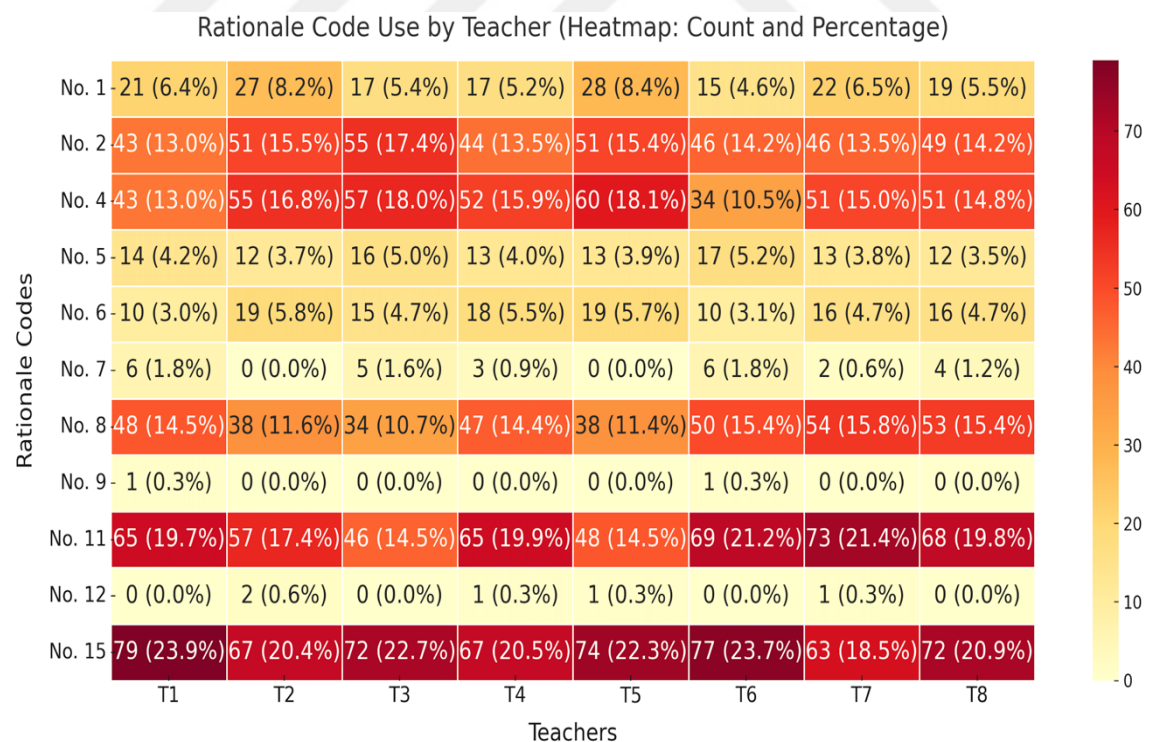
However, despite some subtle differences in their refinement approaches across all papers, overlooked errors still represented the majority of teacher refinements, ranging from 66.5% to 83.7%. This trend underscores a key limitation shown by AI system when responding to student papers, which is overlooking errors while it may perform reasonably well in terms of precision, its lack of comprehensiveness, particularly in identifying all relevant errors, remains a major shortcoming for the accuracy of output and usability of it as classroom feedback.

4.1.2 Analysis of teachers' rationales for the refinements of error-coded feedback

The quantitative insights provided above do not reveal which types of errors the AI tends to miss, but instead offer a quantitative summary of common patterns of teacher-made refinements. To better understand the nature of these issues, this section explores teachers' reasoning behind their modifications quantitatively, as recorded through a rationale codebook presented in a drop-down list format. This codebook made it possible to gain insight into how teachers interpreted the nature, significance, and impact of each error. This data enabled to complement the previous data making teachers' decision-making process explicit, and thus it becomes possible to access into teachers' professional judgment of AI-generated feedback. The following Figure 4.1 presents the frequency distribution of these rationale codes across all teachers in the form of a heat map. The visual scale ranges from red to orange to light yellow, reflecting the relative intensity of code usage by teachers.

Figure 4.1

Frequency and Percentage of Rationale Use by Teacher for Error-Coded Feedback



As can be seen in Figure 4.1., the heat map includes a range of rationale codes reflecting the teachers' underlying reasons for their refinements.

Among the codes provided above, the most frequently employed rationale code across all teachers was No.15 (Accurate and actionable error feedback). This code represents feedback that is correct, clear, and constructive where teachers considered the AI's feedback sufficiently accurate and actionable, requiring little or no intervention. This finding is promising, indicating that AI is capable of useful feedback in certain cases. However, this was closely followed by Codes 11 (Partially accurate error description) and 2 (Major or multiple overlooked errors), highlighting significant limitations in the AI's feedback.

Code 2 was also frequently applied when teachers encountered major or multiple overlooked errors, often involving critical sentence-level issues like fragments, missing subjects, or confused structure resulting in unclear meaning. In these cases, teachers noted that the AI's failure to recognize structural or meaning level problems and provide RW codes to fix them rather than individual codes.

Besides, as another common code, Code 11 was selected often, to indicate error descriptions were only partially accurate, typically due to vague, incomplete, or mixed explanations. This code was also used by teachers when AI correctly marked a sentence in the document but failed to acknowledge this in the accompanying spreadsheet.

Rationale Code 8 (Minor overlooked error) was also frequently used, indicating that while the AI detected some issues correctly, it often missed smaller surface-level errors (i.e. spelling, punctuation etc.) in the same sentence. Though teachers considered code 8 less serious than Code 2 cases, they still required teacher refinements to ensure complete feedback for students.

4.1.3. Qualitative analysis of teacher refinements for AI-generated error-coded feedback

To complement the quantitative findings derived from frequency counts of refinements as well as counts of teachers' rationales, a qualitative analysis of teachers' written review notes for their edits were examined to demonstrate the exact and frequent type of errors refined frequently by teachers. These notes provided concrete examples of what was changed, how teachers suggested those changes, and more importantly, why those modifications were deemed necessary from teachers' pedagogical perspective. Representative quotes that support the findings related to error-coded feedback in RQ1 are provided in Appendix-11.

4.1.3.1. Overlooked rewrite (RW) code

A major issue identified consistently across teacher reviews was the AI's failure to assign the Rewrite (RW) code in sentences requiring complete revision. Teachers typically use this code when individual error codes alone cannot effectively address the issues present.

However, the decision of using RW code involves multiple layers of consideration. Teachers may choose to assign the RW code in cases involving deeper meaning-level issues, structural problems, or due to the presence of multiple errors. In some instances, they believe the RW code should override all other individual codes, but the decision making is not clear-cut, and for a partial part, they can employ RW code while the rest of the sentence they can use individual codes. Teachers may also consider sentence length as another factor; that is, they can assess whether a longer sentence with several errors can still be treated with individual codes or whether the RW code would be more appropriate. Additionally, teachers are also cautious not to overuse the RW code, as this may limit students' opportunities to engage with and learn from specific corrections.

Thus, the decision to use the RW code does not have a clear cut-off point and therefore requires a nuanced understanding and analysis of the student's sentence. That is, the decision involves a multifaceted evaluation and careful judgment regarding the sentence's clarity, structure, length, and overall intelligibility. A common agreement among teachers is to prioritize semantic coherence over grammatical accuracy and to assess whether individual coding would suffice or whether the RW code is an absolute necessity in such cases.

Teachers emphasized that the AI often failed to recognize when sentences needed full rewriting due to:

Sentence fragment problems:

One common issue identified across teacher reviews was the *Overlooked RW code for sentence fragment problems* that required full rewriting, often leading to the absence of the RW code in structurally incomplete or incoherent sentences. Similarly, another sentence beginning with “Since increasing fast food restaurants...” was assigned VF and WW codes, but T2 highlighted that the AI failed to recognize the sentence fragment and the need for a rewrite, stating: “RW code should be added to the first clause so that the student can rewrite it.”

In another instance from Paper 2, teachers noticed RW code was not assigned for fragment sentences. For the sentence “Although we have knowledge about includes in traditional foods” T1 commented: “This is not a full sentence and should be rewritten, but RW code was not provided... Without RW, I do not believe the student can fix the problem.” Likewise, in Paper 4, the sentence “In my opinion 2 causes of the growing role of using robots” was flagged only for word form and spelling. However, T5 explained this as a problem of “Incomplete sentence fragment (‘there are’ missing), there is a need for rewriting,” again highlighting the need for RW code. The same theme also appeared in Paper 5, where the sentence “It is such a problem that mentally or physically” was flagged with a WW code. However, T5 noted: “This is not a full sentence, but there is no feedback for it... WW code was not fixing anything.” Here again, the absence of RW was seen as a serious shortcoming because teachers highlighted unanimously that the individual codes assigned to the sentence were not sufficient to fix the inherent meaning problem, and that the student might be unable to revise the sentence effectively without a more appropriate code like RW.

Clarity-related issues:

Overlooked RW code for word choice related clarity problems also emerged as another issue. Clarity issue appeared for different reasons in students’ paper, and these situations were not provided RW code. For instance, a similar issue appeared for unclear expressions needing RW code for clarification rather than individual codes. In Paper 8, AI provided only punctuation feedback (P) for the sentence: “Especially, McDonald's (P) Domino's (P) Burger King...(P) people need food but it isn't fast food.” T2 noted that this sentence lacked clarity and needed rewriting, and stated, “It isn't fast food needs rewriting. AI should help the student to reconsider what he/she means for clear expression.” These examples show how AI sometimes misses opportunities to support clearer communication, because student likely try to say that fast food is not an healthy option, but AI did not ask for clarification, and because of this, student, when received this feedback can miss opportunity for improving his/her point.

In Paper 5, the AI failed to recognize unclarity in the sentence that was nearly incomprehensible. Instead RW code, the AI assigned multiple surface-level codes to the sentence: “Fast food restaurants be (VF) a barrier (SP) against to (X) them and do not help their self improves (WW).” Despite these codes provided by AI, T5 emphasized that the

real problem was the sentence's lack of clear meaning, stating: "The sentence is not clear in meaning: barred against what? health? How is this helping personal improvement? not meaningful at all." T5's comment highlights that individual error codes were not helpful here, and a full rewrite (RW) was necessary to guide the student in expressing their idea more clearly.

Another clarity-oriented issue overlooked by AI was related to lack of pronoun reference, leading to clarity and coherence problem. As T4 explained for paper 10:

AI tries to clear up basic problems, but it missed the pronoun reference errors and therefore the sentence does not make sense. It is not clear what 'they' refer to. And also it is not clear what 'this situation' refers.

T4 also elaborated on what she might include in an error description, stating: "Then they establish companies: Who is 'they'? The robots? 'This situation getting bigger': What exactly is growing?" Thus, teacher believes that s/he can solve the pronoun reference problem, causing clarity and coherence problem.

In another example, T2 also reasoned why AI does not label clarity issues with RW code. She suggested that the AI's prioritization of grammar over meaning could be the underlying reason for its failure to assign the RW code for such situations. In Paper 8, for the original sentence: "In summary, life is short but your body is very strong and long," T2 commented: "Your body is very strong and long' (RW for clarity)," and elaborated, "The grammar can be correct, but what about meaning?" This reflects a concern that the AI focused too much on surface-level grammatical issues, missing deeper meaning-based problems that required a more holistic and semantic understanding.

Overlooked code for poor thesis structures:

Teachers frequently observed *Overlooked code for poor thesis structures*, that is thesis statements were checked only for grammar, not for structural or logical weaknesses. In paper 3, several teachers also criticized a poorly structured thesis sentence. For instance, one such example came from T1, who commented: "Ss needs to rewrite a better [thesis] statement, so he [/] she needs to change the whole sentence," which suggests the need for full rewrite and the sentence is not fixable through individual error codes. A similar issue emerged in Paper 8, where the sentence "So there are three effects of increasing popularity of fast-food restaurants; be (VF) obeze (SP), spend money and unhealthy (WF)" lacked parallelism and logical coherence. Although the AI applied multiple

individual codes (VF, SP, WF), it overlooked the need for restructuring the sentence. T6 commented:

Thesis needs to have parallel form in verbs, but AI couldn't catch this nuance. I suggest VF code for 'spend.' 'Spend' should be 'spending.' WF code for 'unhealthy' is an unnecessary code.

Likewise, in Paper 7, the sentence "Foods are it can be (WW) wasting and loss effects the (A) decrease (SP) popularity of traditional food restaurants." received individual corrections for word choice, articles, and spelling. However, the teacher found these insufficient and emphasized that the sentence needs rewriting for proper thesis statement:

The sentence has a serious problem about structure, and it is grammatically incorrect, and it is not a proper thesis statement either. RW code was not provided to these cases oftentimes. This can be a nuance it falls short of.

A similar issue also occurred in Paper 8 with the topic sentence, which received some individual codes: "The second effect is spent (VF) money because people need to (X) a lot of money." However, T6 commented, "The topic sentence is vague. Needs rewriting." From this standpoint, it appears that grammatical corrections alone are insufficient to resolve the issue, likely because the topic itself is not clearly stated in the sentence.

Tone problems:

Besides these, *tone problems* in students' written language were not addressed with the RW code in AI feedback, although speaking tone was not deemed appropriate for an academic writing. T4 and T5 both pointed out that the tone issues in students' conclusions were overlooked by the AI. In Paper 5, the student wrote, "I know I mentioned one positive side and two negative sides, but it has not to be all negative or positive every time," which was marked only with a word choice (WW) error, commenting on this, T5 stated, "Tone is not academic and meaning unclear. RW needed."

Similarly, in Paper 10, the student concluded with, "This is a problem as I wrote last paragraphs, and we must solve this problem." In response, T4 remarked, "This is not academic way of summarizing ideas." These comments indicate that the AI failed to detect informal or unclear tone in students' writing.

These findings suggest that the AI prioritizes grammar over meaning, missing opportunities to address deeper structural and communicative issues in student writing

4.1.3.2. Overlooked codes: failure to detect meaning-level problems

During their reviews of AI feedback, teachers also observed that word choice issues were also among the frequently missed errors by the AI-generated feedback. For instance, in Paper 2, the sentence “Maybe fast food is a (A) opportunity for using more time, but most people consume fast food because of (X) it is easy” shows this problem. While the AI correctly identified the article (A) error, it failed to detect two significant word choice problems. As T6 commented,

Two major overlooked mistakes regarding word choice were overlooked, and this can affect understanding of the sentence if not corrected. It should be ‘using time effectively’ and ‘it is easy to order’.

Another instance from Paper 2 further highlights the issue of overlooking deeper problems. In the phrase “how much eating fast food,” the AI flagged a verb form (VF) error. However, teachers pointed out that the major issue was the lack of a clear subject, making the verb form feedback ineffective without addressing the real problem. T2 described this as a major oversight: “Missing subject is a major overlooked problem ... (how much we eat).”

In addition to missing word choice issues, teachers also observed that the AI often failed to identify multiple errors occurring within the same word or phrase. In Paper 5, the sentence “People's Choice (C) affect them and their lifestyle shapes according to their decision” illustrates this pattern. Although the AI marked "Choice" with a (C) code, T7 pointed out that this was insufficient, as the word also involved a singular/plural (S/P) issue. She noted that the correct form should be “choices.”

In line with the issue of the overlooked RW code, the missing feedback on word choices further suggests limitations in the AI's understanding of semantic-level problems. This may be due to the AI accepting a sentence based on its grammatical correctness rather than its meaning.

4.1.3.3. Inaccurate codes: failure to provide appropriate codes for correction

One of the recurrent issues identified across the reviewed papers was the failure of the AI to provide accurate error codes. This typically occurred in two ways: either through the assignment of an incorrect error code or by marking sentences with errors as correct despite noticeable errors.

To start with, inaccurate coding occurs even when sentences with grammar issues are mistakenly marked as correct by the AI. In Paper 6, as observed by T7, the AI marked a sentence with a ✓ tick code despite clear errors: “By trying to make producing and working faster and better, scientists put robots to the job and remove humans from their job.” According to T7, this sentence contains both word form and word choice issues, and she noted: “AI ticked the sentence, but ‘producing’ (WF) should be ‘production.’ And also, ‘put the job’ is not natural, job is WW and should be ‘put to work.’”

Similarly, in Paper 5, AI’s tendency to prioritize grammar over meaning is evident. The sentence “They always eat and eat” was marked as correct by the AI, even though it sounds awkward and unnatural in English. Although the grammar is technically correct, the repetitive structure affects the sentence’s naturalness. As T8 pointed out, “AI marked They always eat and eat’ as correct, but the phrase is awkward and unnatural.” She further explained, “It likely passed because AI focused on grammatical correctness rather than natural usage. The RW code was added for better phrasing.”

Another frequent issue involves inaccurate coding of word choice-related errors, which is also closely related to the prioritization issue mentioned above. For instance, in Paper 4, the AI incorrectly identified “learing” as a spelling (SP) error in the sentence “Because these companys (SP) are makes (VF) new items and learing (SP) to new ideas with people.” However, T2 pointed out that this was a wrong word choice, not simply a spelling mistake. She commented, “Instead of SP, WW code should be given. Major mistake for AI.” This shows that AI failed to recognize the meaning-level problem caused by the incorrect word.

Further evidence of AI’s focus on surface grammar appears in Paper 2. In the sentence “Our metabolism being (VF) slower so we can gaining (VF) weight easily,” the AI marked both “being” and “gaining” with verb form (VF) errors. However, T1 explained that “‘being’ must be changed to ‘get’ slower. Rather than a verb form issue, it is a wrong word choice, and AI did not understand it.” This feedback reveals that the AI tends to check for grammatical structure but often fails to detect inappropriate word choices, leading to meaning problems.

Finally, in Paper 6, additional cases reinforce this pattern. In the sentence “Robots have many mission (S/P) for helping people,” the AI applied an SP (singular/plural) code. However, T2 observed that the feedback was not fully accurate: “As for the SP code, it

doesn't seem accurate. WW would be a better fit, and the word should be 'functions' instead." Another related case involved the word "knowns," where T7 remarked, "'Knowns' is not a word choice problem. VF should be the code. The correct version is 'As it is known,' which is the passive form." This can indicate AI's parsing problem in processing student sentence, caused by not analyzing the sentence as a full clause but rather analyzing words in isolation without focusing on the meaning.

These examples further suggest that the AI struggles to grasp deeper meaning-related issues and often assigns codes that address only surface-level errors, without enabling full correction of the student's intended meaning. Overall, they show that the AI may not fully analyze subject-verb relationships or the meaning of the entire clause, instead parsing words in isolation without adequately considering the sentence as a whole or the intended meaning.

4.1.3.4 Redundant codes: unnecessary error marking

Another pattern observed in Paper 2 was that when the AI failed to recognize a fundamental problem in sentence structure, it sometimes assigned redundant error codes instead. (See table..., section 4 for Redundant code for the representative quotes). For example, in Paper 2, T5 mentioned, "redundant code: WF for 'more'. 'More' is correct.," referring to the sentence "Cooking traditional foods more (WF) difficult than...". The AI failed to detect the missing verb "to be" and did not understand the intended comparison ("Cooking traditional foods [is] more difficult than..."). As a result, not recognizing the missing verb issue led to this redundant coding. This also highlights the AI's difficulty in comprehending both the student's intended sentence structure and their intended meaning of comparison.

Likewise, in Paper 7, T2 stated that an unnecessary WF (word form) code was assigned by the AI for the sentence: "Second (WF) important effect the (WW) waste of food." T2 remarked, "Unnecessary code was given for 'second'," and from the teacher's perspective, the likely corrected version would be "The second important effect is the waste of food." From this perspective, it appears that the AI does not assign codes by considering the full sentence and the meaning to be conveyed, but rather by parsing elements separately.

Besides these, In Paper 8, where the original sentence was: "People should be healthy (WF) body, doing (VF) sport and don't (S/V) eat unhealthy food.", T7 commented the code for healthy is redundant code, stating that: "I added WW code for the word 'be'"

which should be verb: have. The code for healthy is redundant.” This correction highlights how the AI failed to identify the inappropriate verb choice and instead focused on "healthy", which is actually the correct form in that context, and yet provided for this correct word a redundant code. The teacher’s revision demonstrates a deeper understanding of verb–noun collocations, which the AI overlooked and provided redundant code.

4.2 Analysis of Teachers' Refinement for AI-generated Content Feedback

To address RQ1, this section focuses on teachers’ evaluations of AI-generated content feedback, which differs from error-coded feedback in terms of its focus. Unlike error-coded feedback, teachers’ refinements to content feedback aim to improve the organization, development, clarity of ideas, relevance, and overall helpfulness of the feedback for students. Additionally, since the purpose of content feedback is to provide targeted comments on two or three strengths and prioritize three areas needing improvement, teachers also consider issues of prioritization and often suggest changes accordingly when refining AI-generated feedback. The analysis is presented in two parts: first the quantitative findings, followed by the qualitative insights.

4.2.1 Quantitative analysis of teachers’ content feedback refinement patterns

As a part of the quantitative analysis, acceptance levels of teachers and rationale coded selected by teachers were counted to show frequency patterns. First, the analysis of teachers’ acceptance levels was conducted to determine how frequently AI-generated content feedback was accepted, partially accepted, or rejected. Second, the analysis of rationale codes was carried out to better understand the underlying reasons behind teachers’ refinements of AI-provided content feedback.

Table 4.4 presents teachers’ acceptance levels, the rationale codes they selected, and whether these codes were co-selected with others during feedback refinement. This helps to illustrate the key reasons behind teachers’ decisions to refine or approve AI-generated content feedback.

Table 4.4*Acceptance Levels and Rationale Code Analysis for Content Feedback*

Acceptance Level	Rationale Code	Total Count	Percentage	Used Together with Another Code
Partially Accepted	No 19	50	40.7%	Yes (with no 22, 47 times, 65.3%)
Partially Accepted	No 22	47	38.2%	Yes (with no 19, 47 times, 65.3%)
Accepted	No 23	22	17.9%	—
Partially Accepted	No 20	3	2.4%	—
Partially Accepted	No 21	1	0.8%	—

Note: Rationale Code meanings:

- No. 19: Partially accurate content feedback
- No. 22: Inconsistent guidance (mixing guidance with direct corrections)
- No. 23: Fully accepted content feedback
- No. 20 & 21: Other partial acceptance reasons

The table above indicated a notable difference in acceptance levels of content feedback compared to error-coded feedback. The main difference is that content feedback did not receive any rejections from teachers. However, in the distribution of acceptance levels, it can be seen that only 17.9% of content feedback was fully accepted (Code 23), while 82.1% was partially accepted.

When the rationale codes selected by teachers were also examined, the table above shows that the most common rationales for partial acceptance were Code 19 (partially accurate content feedback, 40.7%) and Code 22 (inconsistent guidance, 38.2%). These two codes frequently co-occurred (47 times), suggesting that they represent prevalent concerns for missing or overlooked elements in AI feedback as well as guidance related issues, suggesting the need of more developmental guidance in AI-generated feedback.

4.2.2 Qualitative analysis of teachers' content feedback refinements

In addition to analyzing teacher acceptance levels and selected rationale codes, a content analysis was conducted on teachers' open-ended revision comments. They were examined to identify common patterns in refinement practices. Teachers' comments were

initially coded to detect patterns, followed by a qualitative analysis to interpret recurring themes (See Appendix-12 for the content analysis results and a sample coding, and Appendix-13 for the representative quotes that support the RQ1 findings regarding content feedback.). As summarized in Table 4.5, teachers highlighted both limitations, such as lack of guidance and reflection, and strengths, including useful ideas for content and language development.

Table 4.5

Summary of Content Feedback Issues and Strengths

Theme	Sub-theme	Count	Percentage	Description
Shortcomings of AI Content Feedback (Total = 108; 87.9%)	The need for Content Development Feedback (CDF)	61	49.6%	Feedback lacked sufficient depth, relevance, or specific suggestions to improve content.
	Direct Feedback (DF)	25	20.3%	Feedback was too direct and unguiding, not allowing student reflection.
	Feedback Tone (FT)	13	10.6%	Lacked reflective or guiding tone to improve focus or clarity.
	Missing Feedback – Error (MFE)	4	3.3%	Content feedback mentioned an error, but it was missing in error-coded section.
Strengths of AI Content Feedback (Total = 18; 14.6%)	The strengths of suggestions			
	Good suggestions for supporting ideas	12	9.8%	Feedback included helpful comments for developing ideas.
	Good suggestions for language use	6	4.9%	Feedback correctly highlighted effective or accurate language use.
	Grand Total	123	100.0%	

Most of the teachers' remarks focused on the shortcomings (totaling 87.9% of all comments), with the most frequent theme being Content Development Feedback (CDF), which accounted for 49.6% of all remarks. This was followed by Direct Feedback (DF)

at 20.3% and Feedback Tone (FT) at 10.6%. A smaller number of comments pointed out missing feedback related to content errors (3.3%) and language use (1.6%). In contrast, Positive Feedback (PF) was the only strength-based theme identified and made up 14.6% of the total responses, which include praise from teachers for the suggestions generated by AI for the supporting ideas and language use points.

This analysis highlighted Content Development Feedback (CDF) as the most critical area needing improvement, constituting nearly half of all identified issues. This suggests that AI systematically overlooked opportunities for students to expand, refine, and logically organize their ideas. Therefore, further analysis of the CDF revealed three main sub-categories below:

- **Unaddressed Repetitive Ideas (29.0%):** The AI frequently failed to identify repetition across paragraphs, particularly in body paragraphs, and also in topic and thesis sentences.
- **Missing Structural Feedback (8.9%):** This included overlooked feedback regarding missing conclusions, thesis statements, hook sentences, and cause-effect emphasis.
- **Missed or Insufficient Feedback for Body Paragraph Development (11.4%):** This highlights AI's failure to address expansion issues, offer actionable suggestions, or identify circular reasoning.

To better understand these three CDF categories, both frequency counts and teachers' qualitative comments on their edit notes were analyzed, where they shared their suggested revisions. The analysis below proceeds from the highest to the lowest percentage of content feedback gaps.

4.2.2.1. Unaddressed Repetitive Ideas (CDF Category)

To begin with, the majority of teachers identified *unaddressed repetitive ideas* as a primary concern regarding the usefulness of AI-generated content feedback in their classrooms. Within this theme, several related sub-categories emerged based on teachers' comments, as presented in Table 4.6.

Table 4.6*Unaddressed Repetitive Ideas*

Sub-category	Count	Percentage
Repetition of ideas Not Addressed for body paragraphs	34	27.4%
Repetition of ideas Not Addressed for topic sentences	1	0.8%
No feedback for repetitive words in thesis statement	1	0.8%
Subtotal	36	29.3%

According to Table 4.6, under the broader theme of Content Development Feedback (CDF), *unaddressed repetitive ideas* emerged as the most dominant sub-theme, accounting for 29.3% of all identified issues. Most of these (27.4%) were related to repetition in body paragraphs, while a smaller number involved topic sentences (0.8%) and thesis statements (0.8%).

Repetition of ideas not addressed in body paragraphs was the second most frequently mentioned concern by teachers. For example, Teacher 6 observed in Paper 8: “Obesity and the health issues are basically the same things,” and also expressed concern over the lack of feedback provided on this repetitive content in the body paragraphs, stating, “The student didn’t get feedback about it, which has a serious effect on content.” Similarly, in Paper 6, Teacher 2 stated similarity of body paragraph ideas, stating: “The ideas in the first and second body paragraphs about robots replacing human beings are repetitive, yet no feedback was given for this repetition.” In Paper 5, T 4 added, “I think the subjects for the first and second bodies are nearly identical in student writing” and suggested a solution for the problem: “They can be merged in a single paragraph with concrete examples. The learner must find another aspect to focus [on] for the third body paragraph.” A similar observation was made for (Paper 10) by T4: “This student needs better guidance for body paragraphs 1 and 3. The ideas are quite identical, though the student is trying to present two consequences: one is replacing their jobs, and the other is losing their jobs.”

Additionally, T3 revealed another significant theme which was named *Repetition of ideas Not Addressed for topic sentences*. For paper 10, T3 highlighted repetition at the topic

sentence level unlike T4, who refer to this as problem of repetition of ideas in two body paragraphs. T3 (paper 10) commented,

... Job loss and job displacements paragraph. Actually, the ideas are very similar, especially topic sentences. Maybe changing wordings would help, but AI did not provide any feedback on this.

This observation may indicate that the AI fails to detect repetition in the topic sentences of body paragraphs, which could be used to help students more easily identify and fix repetition problems across paragraphs.

Another issue detected by a teacher was *no feedback for repetitive words in the thesis statement*. T3 (Paper 9) commented: “There seems repetitive words in the thesis statement, e.g., ‘robots’ and ‘people’.” In this regard, the repetition of some words in the word level of the thesis statement were also not noticed by AI.

4.2.2.2 Missed or insufficient feedback for body paragraph development (CDF category)

The last category identified under CDF is *Missed or Insufficient Feedback for Body Paragraph Development*, encompassing 11.4% of total feedback issues. This category covers problems expressed by teachers to explain missing feedback on the structure of the essay for the paragraph development level, which affect students’ organization, development, and logical flow of ideas. Hence, it can be said that this category refers to gaps in higher-order writing support of AI-generated feedback (See Table 4.7).

Table 4.7

Missed or Insufficient Feedback for Body Paragraph Development

Sub-category	Count	Percentage
Expansion problem is not addressed	5	4.1%
Needs More Actionable Suggestions	3	2.4%
Circular reasoning not addressed within paragraphs	2	1.6%
Misinterpretation of student idea and missed feedback for developing student's idea	2	1.6%
Unnecessarily given/redundant feedback for the supporting idea	1	0.8%
Robotic/mechanic feedback lacks in depth for improving the student's paragraphs	1	0.8%
Subtotal	14	11.4%

Among all the sub-categories, *expansion problem not addressed* was the most frequently mentioned issue (4.1%). For instance, T6 (Paper 7) noted:, “The other important problem is ignoring the expansion in the content for body 1 and 2, which makes this essay unbalanced”. Likewise, T1 (Paper 10) highlighted the lack of guidance on paragraph length for fostering expansion problems, stating, “The student should also receive comments on the length of body paragraph 1, as it requires expansion due to a lack of supporting details.”

Another noteworthy theme that emerged was *needs more actionable suggestions*. Three teachers emphasized that the AI provided either vague or insufficiently actionable guidance. T7 observed in (Paper 7) that, “The suggestions for the body paragraphs are quite useful, but maybe could be even more effective with slightly more specific actionable ideas... Student may need a suggestion of an idea like that, otherwise, they may read it, but not apply it.” Likewise, T4 made a similar observation for (Paper 8), and suggested AI feedback could have guided the student simply by asking, “So what, why not fast food?”, noting that ideas were left unfinished and unsupported, and also refers to possible problem in actionability of AI-suggested feedback, “...Wanting from student completely new thing may be hard for student to incorporate into his/ her writing, especially if student does know about the obesity rates.” The reason may be because teacher probably take into consideration that student would write second draft based on feedback in-class, and as it is not allowed to use phone, student may find it hard to incorporate AI’s suggestion.

Two additional sub-categories that emerged were *misinterpretation of student idea* and *missed feedback for developing student ideas*. As T6 (Paper 5) observed, “The content feedback given for the 3rd body paragraph is misleading... the student is not talking about marketing strategies.” The same issue for the same paper was also expressed by T1 as well. This indicates AI provided feedback on the idea present in the essay, but it is not checking for the intended meaning of student because there are times students cannot fully express their ideas and end up in expressions devoid of meanings.

In connection with the same issue, *Circular reasoning not addressed within paragraphs*, was also observed in the two separate teacher reviews of Paper 5. (T3) explained, “People are obese, they eat fast food, they're obese 'cause they eat fast food...”, indicating the

sentences are repetitive in a circular manner not contributing development of content. T2 also expressed the same exact issue for paper 5. This again suggests AI cannot notice if the idea of the student is contributing to content development or expansion of ideas or not.

Additionally, some teachers flagged *unnecessarily given/redundant feedback for supporting ideas*. They argued that the AI commented on content that was already sufficiently developed. To illustrate this problem, T8 (Paper 9) suggested, “AI should be trained to think what the student paper needs... if ideas for body are enough, there is no reason of wanting more.” This shows that AI is not very apt in prioritization of important needs and cater for these real needs of content.

Lastly, in response to AI’s meaning and contextual level understanding problems, another category emerged for AI’s feedback delivery, which is *Robotic/mechanic feedback lacks in depth for improving the student's paragraphs* in AI feedback. One of the teachers (T6, Paper 8) notes, “The content feedback is not enough to make students improve the body paragraphs. The content feedback is so robotic.” This suggests that AI feedback sometimes lacks the insight and flexibility that teachers use to support students' paragraph development. Taken together, these findings highlights that teacher input remains essential in complementing and refining AI-generated content feedback.

4.2.2.3 Missing structural feedback (CDF category)

As for the last sub-category under the main category of CDF, missing feedback for some structural organization of writing emerged based on teachers’ refinement suggestions. These included lack of feedback on the missing conclusions, thesis statements, hook sentence, and absence of cause-effect language in topic/thesis sentences (See Table 4.8)

Table 4.8

Missed or Insufficient Feedback for Body Paragraph Development

Sub-category	Count	Percentage
Missing feedback for the conclusions	7	5.7%
Lack of Thesis Statement not mentioned	2	1.6%
Lack of feedback on thesis/topic sentences for emphasizing cause/effect	1	0.8%
No feedback for the hook sentence missing in the introduction	1	0.8%
Subtotal	11	8.9%

Among these, the most observed problem was the *missing feedback for the conclusions*. For instance, T2 (Paper 8), “Feedback for conclusion should be given to the students because restatement of ideas is missing, which is a major mistake”, and T2 also added, “Final thoughts should be added.” for proper conclusion of the essay. Similarly, in Paper 10, Teacher 7 remarked, “AI did not check whether the conclusion effectively restates the main points of the essay...this issue appears in all the essays so far.” These comments indicate that AI failed to prompt students to reinforce their central ideas in the conclusion, which is critical for effective closure, and something teachers’ pay attention for students to have well- organized essay.

Another thing AI falls short of is *not addressing lack of Thesis Statement* in the essays. To illustrate that, T6 (paper 7) suggested, although thesis was not written, AI just mentioned it as an accuracy issue: “The student doesn't have a problem with the accuracy of thesis statement, there is no thesis statement at all. Similarly, T5 also acknowledged the lack of thesis statement for paper 7, and espouses AI for providing thesis reformulation directly: “Instead of mentioning it and guiding the student to write, thesis statement is given”

Besides, one of the teachers (T7) also noticed *Lack of feedback on thesis/topic sentences for emphasizing the focus of cause/effect* (i.e. positive, negative effects or both). Teacher 7 (Paper 5) reported that s/he was not concerned for the feedback given for the accuracy of thesis, stating: “I think the biggest problem is in the thesis. Yes, it is grammatically correct, but AI should provide feedback to the student to help them categorize their ideas into negative effects and positive effects.” This highlights that teachers’ concern is not the accuracy of thesis, but its lack of clear position, either causes or effects. This suggests that teachers expected AI to provide feedback for the clarity and focus of the thesis.

The same pattern of lack of feedback was also observed for the introduction opening sentence, which is also called the hook sentence. For the missing hook sentence, T3 (Paper 9) commented, “I don't think there's a good hook sentence available, and there's no FB on this.” This suggests that AI did not pay attention to the parts of the introduction, especially when there is a missing element like the hook, or in other words, the attention-grabber sentence.

4.2.2.4. *Direct feedback (DF)*

Aside from the sub-categories that emerged under the main theme of CDF, other independent themes also emerged that are worthwhile to mention. The second major theme identified as an independent category and a shortcoming of AI-generated content feedback is *Direct Feedback (DF)*, comprising 20.3% of the total. Under this theme, teachers expressed twenty-five times that the thesis was provided with a direct, open solution, which contradicts the metalinguistic nature of feedback, as summarized in Table 4.9.

Table 4.9

Direct Feedback (DF)

Issue	Count	Percentage
Direct Thesis Feedback without guidance	25	20.3%
DF Total	25	20.3%

Under this category, all teachers unanimously emphasized that such direct suggestions prevent students' self-correction and thus limit learning opportunities and creative production. T4 (also on Paper 4) warned that, "I don't think that it is a good idea to give a direct thesis statement sentence. The student can copy it directly." Finally, T7 (Paper 1) clearly stated, "Direct thesis statement was provided by AI, and this prevents student from self-correcting their mistake." Besides that, direct thesis statement reformulations are also detrimental to productivity and creativity of students. T4 (Paper 6) noted that "The thesis statement was given directly, which can prevent productivity and originality of the students."

Unlike these comments, one of the teachers, T3 (Paper 4) proposed that he can utilize this thesis reformulation himself, and yet provide students with a part of the thesis formulation to guide them toward better thesis writing, stating with these words: "I could use it after leaving half of the thesis statement blank for students to complete."

These reflections were repeated 25 times throughout the feedback, indicating that AI systematically provides thesis statement directly using a fixed formula. Although teachers did not find the AI-generated thesis format entirely inaccurate, they raised concerns that offering a ready-made thesis limits students' opportunity for reflection and revision.

4.2.2.5. Feedback tone (FT)

The fifth major theme emerged from teachers' suggestions is *Feedback Tone (FT)*, which was expressed by teachers to explain the tone issues in feedback provided by AI. Encompassing 10.6% of all issues, this category covers three major issues, which are lack of reflective and guiding feedback help avoiding repetitive ideas, clarity of papers' focus, and lacks strength and clarity in guiding the student to add supporting ideas. These subcategories can be seen below in Table 4.10.

Table 4.10

Feedback Tone (FT)

Sub-category	Count	Percentage
Lacks reflective/guiding feedback to help avoid repetitive ideas	9	7.3%
Lacks strength and clarity in guiding the student to add supporting ideas	2	1.6%
Lacks reflective/guiding feedback that encourage student to clarify his/her focus	2	1.6%
FT Total	13	10.6%

As the lack of repetitive content mentioned as CDF major issue, so the theme, *lacks reflective/guiding feedback to help avoid repetitive ideas*, reported 9 times in relation to this issue. Teachers pointed out that as AI struggles in identifying idea repetition, it also often fails to guide the student clearly on how to resolve it. T6 (Paper 5) noted that, an easy resolution could have been provided, but not provided by AI, stating: "The student should get content feedback about choosing different ideas for different paragraphs to improve his/her essay." Similarly, T5 (Paper 5) explained another easy thing AI could have mentioned, "It would be too repetitive to mention health issues in both bodies... student was not told that both bodies are similar, and student should not use same points." This comment suggests AI could have commented simply not using same points for developing distinct supporting points for the body paragraph. T7 (Paper 5) echoed the same sentiment and stated his/her practical resolution AI could not address, stating: "To differentiate them, the student needs to explain in Body Paragraph 1 how people become inactive, less energetic, or feel fatigued."

Another issue mentioned by teachers was *lacks strength and clarity in guiding the student to add supporting ideas*. For example, T4 (Paper 2) reported, "...But it could also explain why specific examples are needed, since the student might not think they are really necessary or must be included. I am not sure if the student would take action after reading the feedback." This comment highlights an important concern which is not because AI does not suggest specific examples for improving body paragraphs, but because it does not clearly justify the revision suggestion in a way that encourages the student to address the issue. As a result, the student may not perceive this feedback as a serious flaw.

Lastly, two teachers (T2 and T7) also noted this last sub-category, *lacks guidance encouraging the student to clarify their focus*. For instance, T7 (Paper 5) said, "As for the marketing paragraph, I would ask the student to clarify its focus. I assume this paragraph discusses the positive effects on marketing, but this is not explicitly clear." Likewise, T2 (Paper 8) criticized the lack of prioritization and clarity in support suggestions, stating, "The AI suggested adding more specific examples, but it should have given a stronger suggestion...Otherwise, these paragraphs are not completed from my perspective."

Overall, this theme highlights that even when the AI identifies relevant content issues, it often fails to provide actionable or practical suggestions that guide and encourage the student to resolve them. Additionally, it does not help the student understand how important the problem is.

4.2.2.6. Missing feedback – error (MFE)

Another noteworthy issue, accounting for 3.3% of the total feedback, is *Missing Feedback Error (MFE)*, which refers to instances where errors were addressed in the content feedback but not reflected in the corresponding error-coded annotations, creating inconsistency in the feedback provided (See Table 4.11)

Table 4.11

Missing Feedback – Error (MFE)

Sub-category	Count	Percentage
Error marked in Content feedback, missed in error code feedback	4	3.3%
MFE Total	4	3.3%

Several teachers identified mismatches between error code annotations and written content feedback, particularly in Paper 4. Teacher 1 noted, “In the language use part of content feedback, there is feedback about the verb ‘go’, but there is no error code in the related part in the Word file.” Similarly, Teacher 6 stated, “In line 11, AI marked the sentence as correct, but there is a missing verb problem (give feedback), and this mistake was not included in the error code list.”

Teachers 5 highlighted similar issue for the line 2 of the same paper. Teacher 5 observed, “The mistake in line 2 was corrected in content feedback, but it was not corrected in error code feedback; this shows that AI cannot give the respective code.” This shows that AI feedback cannot give two error codes respectively for the same word.”

These examples show that while some errors are addressed in the content feedback, they are not detected in the error-coded feedback, which undermines the overall consistency and reliability of the AI-generated feedback. This finding can be interpreted as AI’s failure to assign symbolic codes to the errors, despite successfully addressing them in general commentary in an explicit way. This also reveals the general limitations of AI for providing consistently feedback in metalinguistic format.

4.2.2.7. Missing feedback – language use (MF-LU)

In the category of Missing Feedback-Language Use (MF-LU), only two teachers (T7 and T8) mentioned that AI failed to offer feedback related to language use. Although a very small portion, these cases point to suggestions for language use points and cause & effect language use were not perceived sufficient by some of the teachers, as shown in Table 4.12.

Table 4.12

Missing Feedback – Language Use (MF-LU)

Issue	Count	Percentage
Lack of suggestions for Language Use points	1	0.8%
Lack of feedback on the cause & effect language use	1	0.8%
MF-LU Total	2	1.6 %

As seen in table 4.12 above, *the lack of suggestions for language use points* was observed only once. For example, T8 (Paper 5) stated, “I think AI can suggest that the student consider his/her word choices to make their ideas clearer.”

Another problem identified was *the lack of feedback on the use of cause-and-effect language*, also mentioned once. T7 (Paper 5) remarked, "...I would advise the student to use cause-effect language, as I do not see words like 'lead to' or 'cause,' which are essential for clarity." This suggests that the AI failed to provide guidance on language structures that signal cause-and-effect relationships.

4.2.2.8. Positive feedback (PF)

Although much of the teacher commentary focused on the refinement needs of AI-generated content feedback, several teachers also acknowledged the presence of some positive aspects in AI-generated content feedback. This recognition suggests that some teachers perceived the AI-generated feedback as a useful and supportive resource for student writing, including two sub-themes: good suggestions for supporting ideas and good suggestions for language use points. A summary of the percentages and teachers' qualitative comments for each sub-theme is presented in Table 4.13.

Table 4.13

Positive Feedback (PF)

Issue	Count	Percentage
Good suggestions for Language Use points	6	4.9%
Good suggestions for supporting ideas	12	9.8%
PF Total	18	14.6%

The most cited subtheme mentioned 12 times by teachers was *good suggestions for supporting ideas*. Teachers suggested that AI can offer specific, relevant, and helpful content suggestions in line with their expectations and student needs. T7 (Paper 6) noted, "I think AI also asks for specific examples and that is good." Similarly, T4 (Paper 8) found the body paragraph suggestions especially beneficial, stating, "The ideas for the body paragraphs are quite useful and very similar to what I would give... especially comment for body 2 and 3 is straight to the point."

T8 (Paper 7) echoed this view by stating, "...the suggestions for body paragraphs are actually useful, especially for body 3, it makes clear that student need to write more to support his/her idea. I thought the same thing for this body." Another example came from T8 (Paper 6), who found the suggestions relevant and targeted: "The suggestions for

bodies are all relevant and to the point... as AI noted, in the first place, the situation could have been explained better.”

In addition, *positive comments on language use feedback* were noted six times. Teachers acknowledged the usefulness of AI’s metalinguistic guidance for grammar and vocabulary. For instance, T4 (Paper 2) commented, “The comments for language section for the grammar points and suggestion for varied vocabulary are useful.”

Likewise, T7 in Paper 7 emphasized that, “the suggestions for word form and verb form is very useful, as this way, student can understand repeated problems in the essay.” T8 (Paper 6) also appreciated the AI’s vocabulary suggestions, adding, “The repeated mistakes of word form of health were also mentioned... offering students different vocabulary to use instead of fast food is [a] good idea.”

In this vein, despite shortcomings of AI-generated feedback in detection of nuances, AI’s provision of specific examples and suggestions for language points were found useful by some of the teachers.

4.3. Findings related to Research Question 2

The second research question focused on how teachers evaluated the accuracy of AI-generated feedback on student papers, including both error-coded and content-related feedback. Data gathered from acceptance ratings in the structured review and holistic rubric evaluations were analyzed quantitatively, while teachers’ interview responses were analyzed qualitatively to provide additional support. This section first presents the quantitative results, followed by the qualitative findings that complement them for better interpretation (See Appendix-14 for sample acceptance ratings, accuracy evaluations, and reliability analysis results; and Tables 5 and 6 for representative interview quotes supporting RQ2 findings).

4.3.1 Quantitative findings related to AI-generated feedback accuracy for both error-coded and content-related feedback

This section presents first a comprehensive quantitative analysis of how teachers evaluated the accuracy of AI-generated error-coded feedback. The data were obtained from teachers’ acceptance ratings documented in structured review table during teachers’ review process. The table below presents the acceptance level of teachers for error-coded feedback, which enable understanding teachers’ accuracy evaluations (See Table 4.14)

Table 4.14*Acceptance Levels Across Teachers for Error Feedback*

Status	Count	Percentage
Accepted	578	33.4%
Partially Accepted	448	25.9%
Rejected	702	40.6%
Total	1,728	100%

These results in the table above reveal that less than half of the AI-generated error-coded feedback (33.4%) was considered fully acceptable by teachers, while a substantial portion (40.6%) was rejected, and notably 25.9% were partially accepted by teachers. This highlights significant limitations in the AI's ability to provide accurate error-coded feedback.

Considering that this can vary among teachers based on their teacher identity and pedagogical reasoning, teachers' acceptance levels across papers from 2 to 10 were also counted and frequencies were taken (see Table 4.15).

Table 4.15*Acceptance Levels Across Teachers for Error-coded Feedback*

Teacher	Accepted (Count)	Accepted (%)	Partially Accepted (Count)	Partially Accepted (%)	Rejected (Count)	Rejected (%)
T1	81	37.5%	65	30.1%	70	32.4%
T2	70	32.4%	51	23.6%	95	44.0%
T3	73	33.8%	50	23.1%	93	43.06%
T4	68	31.5%	53	24.5%	95	44.0%
T5	74	34.3%	45	20.8%	97	44.9%
T6	76	35.2%	63	29.2%	77	35.6%
T7	64	29.6%	62	28.7%	90	41.7%
T8	72	33.3%	59	27.3%	85	39.4%
Total	578	33.4%	448	25.9%	702	40.6%

The results showed that most teachers tend to either rejected (40,06%) or accepted (33.3%) the AI-generated feedback. The partially accepted feedback represents approximately 25.9% of the total feedback instances. This distribution suggests that AI feedback is either fully accepted or largely rejected by most teachers. This consistency in teachers' acceptance levels suggests a reliable evaluation standards and consensus among the teachers with regard to accuracy of AI-generated feedback.

However, small variances also appeared amongst teachers. For instance, teachers like T2 and T4 have a high percentage of rejected feedback (44.0%), indicating that they find the AI feedback less useful or relevant for their teaching practice. On the other hand, T1 shows a more balanced evaluation, with nearly equal proportions of accepted and partially accepted feedback, suggesting that T1 may be more flexible or open to AI suggestions but still prefers adjustments.

This variation from teacher to teacher suggests that feedback practices and the way AI feedback is read or interpreted may differ based on the teacher's methodology and approach to feedback. Some teachers may be stricter in their evaluation of AI feedback, while others may integrate AI feedback more readily into their own practices. That is, a teacher's individual teaching style, philosophy, and methods of providing feedback also affect how they engage with AI-generated feedback and refine it.

Just as variation was observed across teachers' acceptance levels of feedback, paper-based acceptance levels of teachers also showed considerable variation, suggesting that student error types or the paper quality impact the accuracy of AI-generated feedback (see Table 4.16).

Table 4.16

Acceptance Levels Across Papers for Error-coded Feedback

Paper	Accepted	Partially Accepted	Rejected	Total
Paper 2	64 (33.3%)	78 (40.6%)	50 (26.0%)	192
Paper 3	76 (41.3%)	34 (18.5%)	74 (40.2%)	184
Paper 4	26 (15.5%)	40 (23.8%)	102 (60.7%)	168
Paper 5	58 (24.2%)	111 (46.3%)	71 (29.6%)	240
Paper 6	74 (44.0%)	45 (26.8%)	49 (29.2%)	168

Table 4.16(continued)*Acceptance Levels Across Papers for Error-coded Feedback*

Paper	Accepted	Partially Accepted	Rejected	Total
Paper 7	58 (38.2%)	21 (13.8%)	73 (48.0%)	152
Paper 8	35 (21.9%)	28 (17.5%)	97 (60.6%)	160
Paper 9	114 (44.5%)	51 (19.9%)	90 (35.2%)	256
Paper 10	73 (35.1%)	37 (17.8%)	98 (46.6%)	208

Papers 6 and 9 had the highest acceptance rates (44%), while Papers 4 and 8 had very low acceptance rates (15.5% and 21.9%) and high rejection rates (over 60%). This suggests that the AI's performance varied considerably depending on the nature and complexity of the student writing.

4.3.1.1 Teacher acceptance levels for content feedback

The analysis of teacher acceptance levels for content feedback suggests that teachers do not view AI-generated content feedback as entirely ineffective or unhelpful. In contrast to error-coded feedback, content feedback received no outright rejections, although full acceptance remained limited. This indicates that teachers perceive its accuracy more positively compared to error-coded feedback, as presented in Table 4.17 below.

Table 4.17*A General Overview of Teachers' Acceptance Levels for Content Feedback*

Acceptance Level	Count	Percentage
Accepted	22	30.6%
Partially- accepted	50	69.4%
Total	72	100%

The high rate of partial acceptance (69.4%) indicates that while teachers found content feedback more usable than error feedback, they still believed substantial modifications were needed to enhance its usefulness for students. Analysis by paper revealed considerable variation in content feedback quality, with acceptance rates ranging from 0% (Papers 3, 5, and 10) to 87.5% (Paper 2), suggesting inconsistent performance across different writing samples.

When each individual teacher's evaluation of content feedback was examined, it was observed that all teachers predominantly favored partial acceptance of AI-generated feedback. Their approaches to content feedback showed no major variation, whereas acceptance levels for error-coded feedback varied more significantly due to the greater degree of teacher interpretation involved, as displayed in Table 4.18.

Table 4.18

Acceptance Levels for Content Feedback Across Teachers

Teachers	Accepted	Partially Accepted	% Accepted	Partially Accepted%
T1	2	7	22.2%	77,8%
T2	3	6	33.3%	66,7%
T3	3	6	33.3%	66,7%
T4	3	6	33.3%	66,7%
T5	3	6	33.3%	66,7%
T6	2	7	22.2%	77,8%
T7	3	6	33.3%	66,7%
T8	3	6	33.3%	66,7%
Total	22	50	30.6%	69,4%

Table 4.18 indicates that only 30.6% of AI-generated content feedback was fully accepted by teachers, highlighting a strong need for refinement. In other words, around 70% of teachers partially accepted the content feedback. This partial acceptance suggests that teachers unanimously recognized missing elements in the feedback's completeness and nuanced understanding, which likely resulted from AI's failure to prioritize the issues it should have.

The acceptance levels by papers also showed variability in the AI's ability to detect content-related problems in student writing. However, when the results of content feedback are compared with those of error feedback, the acceptance rates still portrays a more positive outlook on the usefulness of content feedback. The findings for content feedback acceptance levels across papers can be seen in below in Table 4.19.

Table 4.19*Acceptance Levels by Paper for Content Feedback*

Paper ID	Accepted	Partially Accepted	% Accepted	% Partially Accepted
Paper 2	7	1	87.5%	12.5%
Paper 3	0	8	0%	100%
Paper 4	1	7	12.5%	87.5%
Paper 5	0	8	0%	100%
Paper 6	3	5	37.5%	62.5%
Paper 7	3	5	37.5%	62.5%
Paper 8	3	5	37.5%	62.5%
Paper 9	5	3	62.5%	37.5%
Paper 10	0	8	0%	100%
Total	22	50	30.6%	69.4%

Taken together, this section suggested that teachers' acceptance varied across different feedback types. However, these findings do not capture teachers' broader evaluations of AI feedback quality. Hence, the following section presents teachers' holistic evaluations of AI-generated feedback, supported with teachers' rubric commentaries.

4.3.1.2 Analysis of rubric evaluations and comments of teachers on the accuracy of AI-generated feedback

In addition to reviewing individual feedback instances and assigning acceptance levels for each, teachers also evaluated the overall quality of AI-generated feedback using a holistic 3-point rubric, where a score of 1 indicated low accuracy, 2 indicated mid accuracy, and 3 indicated high accuracy. These quantitative ratings were further supported by teachers' qualitative rubric commentaries, which provided insight into the reasoning behind their scores. Table 4.20 presents a summary of these evaluations across ten student papers.

Table 4.20*Rubric Ratings Across Papers*

Paper	Rating 1 (Low)	Rating 2 (Mid)	Rating 3 (High)	Consensus Rating
Paper 2	0	8	0	2 (Mid)
Paper 3	0	8	0	2 (Mid)
Paper 4	7	1	0	1 (Low)
Paper 5	5	3	0	1 (Low)
Paper 6	0	0	8	3 (High)
Paper 7	3	5	0	2 (Mid)
Paper 8	7	1	0	1 (Low)
Paper 9	0	8	0	2 (Mid)
Paper 10	0	8	0	2 (Mid)

Based on the rubric evaluations, only Paper 6 received unanimously high accuracy ratings from all teachers. T3 explained this rating as “Minor errors, overall accurate feedback,” while T7 noted that it was “really helpful feedback with many correct and useful points, but the paper itself does not have complex problems and is written well.” This suggests that the AI’s success may have been partly due to the paper’s relative simplicity. This point was also echoed by Teacher 8, who stated, “Maybe AI found it easy to manage that paper.”

In contrast, Papers 4 and 8 received low ratings from seven teachers. For Paper 4, T5 explained, “The rewrite code was not provided, even though it was necessary for many sentences with structural and meaning-related issues.” For Paper 8, T3 cited, “Missing lines led to rejection, especially lack of constructive feedback.” Similarly, T8 reasoned those various factors contributed to the rejection of Paper 8, stating:

...No feedback for wrong modal usages, no feedback for wrong subject-verb agreement (e.g., subject should be in plural form), or no RW feedback for sentences with expression problems.

Teacher 2 speculated that these extensive problems might indicate, “AI has processing problems, I guess.”

Some of the mid-accuracy rated papers received criticism related to pedagogical issues rather than the accuracy of error or content feedback. For instance, T6 noted that in Paper 2, “Some suggestions are so direct, so such feedback will not help students think and fix their own mistakes,” highlighting concerns about overly directive, non-guiding feedback that prevents student self-correction. Similarly, for Paper 3, the same teacher observed, “A student cannot fix the problems only with the given feedback in some cases,” emphasizing the need for actionable, constructive feedback and suggesting that isolated corrections may not suffice in cases requiring sentence rewriting.

Sharing a different teacher reasoning than the other teachers, T4 explained that Paper 10 received mid-accuracy point because of undetected similar ideas across body paragraphs, stating: “I also partially accepted content feedback because there were similar ideas, and this was not provided any feedback.”

4.3.2 Qualitative findings related to AI-generated feedback accuracy from the teacher interviews

While the count of acceptance levels, rubric evaluations, and teacher commentaries provided useful insights, they were not sufficient on their own to capture the full depth of teachers’ reasoning regarding the accuracy of AI-generated feedback. For this reason, semi-structured interviews conducted with all teachers individually. These interviews were analyzed using content analysis to provide a more contextualized and detailed understanding of the underlying reasons behind teachers’ evaluations. This allowed for the identification of recurring concerns and expectations from accuracy that were not made evident through quantitative data.

4.3.2.1. Qualitative findings for error-coded feedback accuracy

This sub-section dwells specifically on teachers’ views regarding the accuracy of error-coded feedback. The analysis of interviews with all eight teachers revealed three main themes that frequently emerged across their responses: overlooked errors, incorrect or inconsistent code application, and underlying reasons for these recurring issues in AI-generated feedback. *Overlooked errors* refer to instances where the AI failed to detect and mark mistakes. *Incorrect or inconsistent coding* covers both mislabelled errors and the AI’s inability to provide more than one error code for multiple existing issues within the same sentence or phrase. The third theme focuses on the possible causes behind these mistakes, based on what teachers observed and attempted to explain about the AI’s

performance. These themes are summarized along with their sub-codes and the number of teacher mentions in Table 4.21 below:

Table 4.21

Error Feedback Accuracy Issues

Sub-themes	Sub-codes	Total Mentions (n=8)	Teacher IDs
Overlooked errors	Code 1: Missing Rewrite code	8	T1, T6, T4, T8, T3, T7
	Code 2: Missing auxiliary verbs	4	T1, T7, T3, T6
	Code 3: Missing main verbs	3	T1, T6, T4
	Code 4: Missing punctuation	3	T7, T5, T6
	Code 5: Missing prepositions/ articles	2	T6, T7
	Code 6: Missing pronouns	1	T7
	Code 7: Parallel structure	1	T5
	Code 8: Passive/Active structure	1	T6
Incorrect/inconsistent code application		6	T6, T8, T3, T2, T7, T5
Reasons for frequent mistakes in AI-generated error-coded feedback		6	T6 (2x), T4, T8, T2, T1
	Too much focus on grammar	2	T8 and T6
	Not understanding meaning issues	1	T2
	Focusing on errors one by one in isolation	1	T4
	Performance-variation depending on paper	2	T1 and T6

As shown in Table above, *Overlooked RW code* emerged as the most significant theme, mentioned by all eight participants (n=8), with eight distinct sub-codes representing specific types of overlooked errors.

Within this theme, the most frequently mentioned sub-code was *Missing Rewrite code* (Code 1), identified by six teachers (T1, T3, T4, T6, T7, and T8). T1 explicitly stated: “The AI would often miss when a complete rewrite was needed.” Similarly, T2 emphasized: “There were so many overlooked errors, first of all, mainly RW codes.” T4 highlighted the critical importance of rewriting, noting: “For me, the most important one actually was rewriting sections.” These observations indicate that teachers prioritize holistic assessment of writing quality, whereas the AI appears to focus primarily on discrete grammatical elements. However, T8 explained the need to consider meaning: “Some sentences needed rewriting, not because of grammar mistakes,” highlighting that rewriting recommendations often stem from concerns about clarity and expression rather than technical errors.

In this vein, participants provided specific reasons to explain why they added RW (rewrite code). T1 detailed his reason to use RW with these words: “I had to write 'rewrite' frequently because sentences either lacked meaning, were unclear, incomplete, or had some structural problems.” T6 specifically noted the AI's failure to recommend for meaning-level problems: “AI did not provide the RW code even when a sentence is unclear or doesn't make sense.”

T3 explained his use of the RW code, and different from others he extended the list of his usage, “I had to add overlooked codes and use the RW code mostly for run-on sentences, incomplete ideas, repetitive structures, or tone issues.” These direct quotations collectively demonstrate that while the AI can identify grammatical errors, it struggles to evaluate writing in terms of overall coherence, clarity, and communicative effectiveness.

Following this, the most frequent sub-code within *Overlooked errors* was *Missing feedback on verbs auxiliary verbs* (Code 2) reported by three teachers (T1, T3, T7 and T6) and missing main verbs (code 3) reported by (T1, T6, and T4). T1 commented her observation for both missing feedback for auxiliary and main verbs “The most frequent corrections were missing main verbs and auxiliary verbs.” T3 provided more specific detail: “[AI] didn't catch missing auxiliary verbs (am, is, are, was, were) or modals.” As

for *Missing main verbs* (Code 3), T4 also specifically reported this issue: “no feedback for missing verbs...”

The table 4.21 further indicates that teachers each identified some mechanical errors not captured by AI for feedback. Among these, three of them (T7, T6 and T5) mentioned *Missing punctuation* (Code 4). Two of them (T6 and T7) observed missing *Missing prepositions/articles* (Code 5). To illustrate, T6 for instance noted lack of feedback on punctuation: “some missed punctuation errors I can remember, and some commas before conjunctions.”

T7 also stated there was missing feedback for prepositions/ articles: “I often had to add missing codes for articles, prepositions. In addition to these, T7 (n=1) also specifically noted missing feedback for pronoun reference (*Missing pronouns*, Code 6). T7 commented, “I often had to add missing codes for pronouns and pronoun reference errors.” This insight suggests that AI may not be able to trace which pronoun is needed in the sentence, possibly because it reads sentences individually rather than analyzing them in relation to one another.

In addition to the surface-level mechanical errors that were missed, some teachers also pointed out that the AI overlooked important issues related to sentence structure, specifically, parallel structure and passive/active voice (Codes 7 and 8) were mentioned as areas the AI failed to address. For example, T5 mentioned that the AI didn’t give any feedback on parallel structure in sentences, especially in lists of causes or effects. “There was a lack of feedback on parallel structure in the lists of effects or causes, which I always have to correct manually”.

T6 also pointed out a similar issue with passive and active voice, saying: “It also struggled with passive and active sentence structures” (*Passive/Active structure*, Code 8). These examples pinpoint that the AI sometimes failed to recognize deeper problems in sentence construction, suggesting it still struggles to fully grasp the meaning and more complex sentence forms.

Incorrect/Inconsistent Code Application

The second major theme, *Incorrect/inconsistent code application*, was mentioned by six of the eight teachers (T2, T3, T5, T6, T7, and T8) as shown in Table 4.21. This theme encompasses instances where the AI did identify an error but assigned an incorrect code

or inconsistently applied codes across similar errors. T8 observed: "It sometimes flagged errors for the wrong reasons." T2 noted inconsistency: "It might mark S/V, SP, or WF correctly in one sentence but miss it in another sentence."

T6 also added incorrect coding happened for word-choice issues, and stated, "AI provided redundant or incorrect codes, especially for word choice issues," and further explained, "It is unclear to the AI whether the issue is word form or incorrect word choice." T3 provided a specific example: "AI marked a word form (WF) code when it was actually a word choice (WW) problem." He elaborated with a concrete instance: "A student wrote 'traditional fast-food restaurants' when they actually meant 'traditional local restaurants,' but AI couldn't detect the word choice mistake." Similarly, T8 noted: "Not recognizing that 'job lives' should be 'work lives'," All of these instances suggest that grammar-level issues may take priority in the AI's detection of errors. As a result, meaning-level problems are often disregarded, which in turn leads to incorrect coding and incomplete corrections.

Participants also identified the AI's inability to assign multiple codes to a single word or phrase as a limitation within this theme. T7 explicitly stated: "AI cannot provide two codes at once for multiple issues in a word," providing an example: "One clear example was with the sentence about 'growing up' in engineering. The AI provided a VF code but missed that 'growing up' was also the wrong word choice." This suggests AI cannot detect errors when it is dense and at the meaning level.

Reasons for Frequent Mistakes in AI-generated Error-coded Feedback

As indicated in Table X, the third theme, *Reasons for frequent mistakes in AI error-coded feedback*, was identified by five of the eight participants (T1, T4, T6, T7, and T8). This theme addresses the underlying causes teachers perceived for the AI's limitations, with four distinct sub-categories emerging from the data.

One of the most frequently mentioned reasons, *Too much focus on grammar*, was identified by two teachers (T8 and T6). T8 directly articulated this issue: "I believe the AI focuses too much on grammar and doesn't always consider the meaning." T6 similarly emphasized this concern, noting "meaning is not its first focus, but it should be." These two teachers highlight that excessive focus on grammatical structures might be the reason why AI could not provide meaningful feedback on content and clarity.

Another identified reason was *Not understanding meaning issues*, which was specifically mentioned by one teacher (T2). T2 explained: “AI couldn’t detect that the sentence has meaning problems”. This limitation suggests that the system fails to understand meaning of a sentence as uttered by its interlocutor.

T4 highlighted *Focusing on errors one by one in isolation* as a significant limitation: “AI just tried to focus on them one by one, but it can be confusing.” This observation suggests that reading sentences focusing on separate words rather than with sentence-level focus prevents the AI from seeing clarity or coherence related issues.

The final sub-category, *Performance-variation depending on paper*, was reported by two teachers (T1 and T6). Both T1 and T6 noted inconsistencies in AI performance across different student papers, while T1 remarked, “when it comes to weak students, I think AI [pause] most of the feedback was problematic. Good papers, of course, they did not want so much feedback or details, so they were... they were ok.” Similarly, T6 stated, “...I think... depending on the students’ paper, the dynamic of the student’s paper may change. If it is a weak paper, for example, AI can be confused about it.” This indicates that AI-generated feedback is inadequate for effectively supporting struggling writers whose writing include complex structural and meaning related errors.

These findings are really worthwhile because they overall indicate that while AI error feedback provides some value in identifying certain types of grammatical errors, its approach to feedback exhibits certain shortcomings such as; overemphasis on grammar, difficulty of understanding student’s intended meaning, isolated error analysis, and performance variation depending on density of errors. This suggests, AI is not yet ready to do comprehensive error analysis and highlights the need for more semantically- aware AI feedback systems that can evaluate writing holistically.

4.3.2.2. Qualitative findings for content feedback accuracy

With regards to accuracy of content feedback, several themes emerged from interview data that summarizes teachers’ perception of accuracy in more detail and substantiate teacher evaluations of content feedback with qualitative insights from teachers. These findings highlighted several key content feedback gaps observed by teachers communicated during review suggestions. These gaps observed by teachers are summarized in Table 4.22.

Table 4.22*Content Feedback Gaps*

Sub-theme	Total Mentions (n=8)	Teacher IDs
Failure to identify repetitive ideas across paragraphs	8	T1, T6, T4, T8, T3, T2, T7, T5
Missing feedback on short or underdeveloped conclusions	5	T1, T8, T3, T2, T5
Need for more specific, actionable suggestions	4	T8, T2, T7, T5
Misinterpretation of student meaning	2	T6, T8
Providing direct corrections for thesis statements	1	T4

The analysis of teachers' perceptions of AI-generated content feedback revealed several shortcomings. One frequently mentioned problem was *the failure to identify repetitive ideas across paragraphs*, noted by eight teachers (T1, T6, T4, T8, T3, T2, T7, and T5). All these teachers indicated that AI frequently overlooked repeated ideas in the body paragraphs of students' essays, limiting opportunities for improvement. For instance, one teacher stated, "The body paragraph, for example, did not give three different ideas, and AI did not see it." Another observed, "Paragraphs 1 and 2 kept repeating the same ideas...but AI ignored the repetition and didn't suggest how to fix them." These findings suggest that AI's failure to capture similar ideas and its consistent overlooking of such repetitions indicate that it may be processing texts paragraph by paragraph, rather than conducting a holistic analysis and identifying overlapping ideas across them.

A second major issue involved *missing feedback for conclusions*, particularly when they were short or underdeveloped. This concern was raised by five teachers (T1, T8, T3, T2, and T5). One teacher noted, "There was a conclusion that was too short, and there was no feedback for it," while another observed, "It does not provide feedback on the conclusion of the essay throughout the feedback although there was a need for clear restatement." These comments suggest that the AI feedback system lacked adequate attention to the structural requirements of essay writing, particularly in evaluating and supporting effective conclusion writing.

Another recurring theme was *the need for more specific and actionable suggestions*, emphasized by four teachers (T8, T2, T7, and T5). Teachers pointed out that while AI sometimes identified areas for improvement, it often failed to provide sufficiently

concrete guidance. One teacher noted that “AI should also provide revision options, like combining similar paragraphs or helping students develop more distinct supporting details,” while another stressed the need for more detailed suggestions, explaining, “If the suggestions for body paragraphs included more specific, actionable ideas...” Another teacher highlighted that although AI frequently recommended expansion of ideas, it often failed to specify “where [students] can add and build on their ideas.” This feedback illustrates the importance teachers place on not just identifying weaknesses but also offering students clear pathways for revision.

Additionally, two teachers (T6 and T8) reported concerns regarding *the misinterpretation of student meaning*. T6 stated, “AI misinterpreted this as being about marketing strategies, completely missing the student's actual points,” and another teacher echoed a similar experience. These comments point to limitations in AI’s ability to comprehend student ideas when they are not clearly expressed. Therefore, AI’s comprehension of student meaning is not as strong as that of a teacher, who can more easily recognize the intended message.

Finally, only one teacher (T4) commented on AI’s *providing direct corrections for thesis statements*. The teacher explained, “I didn’t like some ideas in content feedback because sometimes AI provided direct corrections, like direct examples. For the thesis, I might want my student to write the thesis statement,” highlighting a pedagogical concern about AI’s tendency to provide direct solutions rather than guide student development. However, this was mentioned only one of teachers, suggesting this was not seen as major problem for the accuracy of AI-generated feedback.

Overall, these key findings from interviews related to accuracy of content feedback generated by AI suggest that it demonstrates considerable limitations in recognizing deeper issues, offering actionable guidance, and accurately interpreting student meaning to help student develop their content.

4.4 Findings related to Research Question 3

The third research question aimed to provide qualitative analysis of the teachers’ interviews for the exploration of advantages and challenges they experienced in refining AI-generated feedback, their perspectives on its future classroom use, and suggestions for its usability in classrooms. The analysis is organized into four distinct sections, each

addressing one of these key aspects (See Appendix-15 for the qualitative analysis of Teachers' interviews for RQ3).

4.4.1 Challenges of refining AI-generated feedback from teachers' perspectives

After examining teachers' evaluations of AI-generated feedback accuracy (RQ2) through excel acceptance levels, rubric evaluations and interview responses related to accuracy of AI feedback, this section addresses RQ3 by exploring the advantages and challenges teachers experienced during the refinement process, as well as their perspectives on the potential future use of AI feedback in classroom contexts. The table below presents challenges and positive aspects of refining AI feedback in Table 4.23.

Table 4.23

Pros and Cons of Refining AI feedback

Sub-themes	Sub-codes	Total Mentions (n:8)	Teachers' Mentionings
Need for substantial teacher refinement		4	T8, T2, T3, T6
Challenges of Refining AI Feedback	Increased teacher time investment for refining AI feedback	4	T7, T8, T6, T2
	Increased teacher cognitive load due to interpretation of AI feedback	3	T1, T4, T5
	Lacks in human judgment/nuance	5	T4, T8, T3, T2, T7
Positive Aspects of Refining AI feedback			
	Easier refinement of content feedback compared to error feedback	5	T6, T8, T2, T7, T5
	Appreciation of AI's guiding tone in error descriptions	1	T4
	Promotion of teacher self-reflection for comprehensive feedback	1	T5

With regard to the pros and cons of AI-generated feedback, all eight teachers expressed AI feedback's *need for substantial teacher refinement* as major theme. Although error feedback was seen as somewhat useful, none of the teachers expressed strong trust in AI-generated grammar feedback for the time being in its current form. T3 reasoned his substantial refinement as the accuracy problem in the AI feedback, and stated, "AI still requires too much extra effort to refine its feedback." T8 expressed his mistrust in AI

feedback and, hence, his substantial effort to refine it to increase reliability: “But to be honest, I had to double-check its suggestions because they weren't very accurate.” T2 also added notes for her substantial revision for the overlooked errors, and inaccurate coding: “AI made my work harder because it missed so many errors, and I had to add them. And also, there were some wrong errors that were provided.” What is more, T6 also mentioned her evolutions in rubric, “Out of all the papers I reviewed, I could only accept one of them. The system has serious problems, maybe the system is not ready ...”. This suggests that the trust towards AI error-coded feedback is very low due to unreadiness of the system for classroom feedback.

Within this sub-theme, different sub-codes emerged from the analysis to express distinct challenges of refining AI feedback. For example, *Increased teacher time investment for refining AI feedback* was also mentioned as sub-category for the substantial revision need by the majority of teachers. T7 commented that rather than fixing student errors, fixing AI's mistakes became the main challenge: “I feel like I spent a lot of time fixing AI mistakes.” T2 pinpointed overlooked feedback as the main reason why she finds it challenging to refine AI feedback, saying that, “AI made my work harder because it missed so many errors, and I had to add them.” T6 further highlighted the degree of intervention required, suggesting that she spent even more time on feedback than she normally would: “I spent so much time, maybe more than my usual practice, correcting them.” This suggests she corrected student errors as well as AI-generated feedback for the student paper.

Aside from the issue of time investment, three of the teachers also described this workload under the sub-category *Increased teacher cognitive load due to interpretation of AI feedback*, often caused by a mismatch between AI and teacher interpretations. This was perceived as both an additional burden and a source of mental effort, stemming from the need to interpret and understand AI-generated feedback in order to refine it. As T4 explained, “This difference in interpretation makes me think a lot and put in extra effort. And it can double the work sometimes for the teacher, I guess.” Similarly, T5 echoed this sentiment, noting, “I had to spend extra effort trying to interpret the feedback, first of all, because my interpretation and AI's interpretation don't match all the time.” T1 also highlighted the difficulty of interpretation, which stemmed from the lack of clarity in the AI feedback, as teachers often have to interpret the reasoning behind it: “I really had

difficulty while I was trying to understand some of the feedback." This difficulty may stem from the possibility that AI-generated feedback includes suggestions or perspectives that teachers might not consider relevant or necessary at the time, or she may point to lack of clarity in the feedback provided by AI.

Another commonly reported shortcoming of AI feedback, as reported by five teachers, was *the lack of human judgment and nuanced understanding*. T4 emphasized the emotional judgement distinction between teachers and AI by saying, "I always consider whether the error codes would be helpful and understandable for the student because I have emotions, and I am a human being. I think for my student as well, but AI doesn't, I think". T2 similarly highlighted structural complexity of a sentence that can be barrier to AI's understanding of student paper: "Maybe it struggles depending on how the sentence is structured, or it doesn't interpret the student sentences well." She referred to the issue of not interpreting students' intended meaning well and noted that AI's interpretation ability varies depending on sentence complexity. T7, confirming this view, also listed some specific dimensions AI misses: "AI needs better detection for meaning level, word choice, content, context, academic tone, and should read and analyze whether the paragraph sentences flow logically." T8 also suggested that AI has limited interpretive understanding when it comes to nuances that are easily grasped by humans but difficult for AI: "It struggles with nuance that a teacher can easily understand." T3 echoed this concern about AI's revision skills, noting that it lacks human-like reading and revision abilities: "Sometimes, a sentence didn't need small fixes. It needed a complete rewrite, for instance, but AI didn't recognize that nuance." From his perspective, AI missed sentences that required rewriting because it does not read papers the way a teacher does and lacks the teacher's nuanced and varied approach to correcting errors.

4.4.2 Positive aspects of refining AI-generated feedback from teachers' perspectives

Despite the negative outlook on AI's error feedback, seven teachers highlighted another important sub-theme indicating *Easier refinement of content feedback* compared to error feedback. These teachers compared error feedback with content feedback and expressed favorable views for refining content feedback to use in their classrooms. T8 compared error code feedback to content feedback and commented: "I found it more useful for content feedback," while T2 noted, "The error code part made my job harder. I think content feedback made it easier. And the feedback overall was useful for the student",

suggesting that the refinement process of content feedback was smoother, and the feedback generated by AI is beneficial for the students. T5 also appreciated the structured and helpful nature of the feedback, describing it as effective with minor modifications: “The feedback is structured and really helpful with some small adjustments that I have made.” The only participants who favors content feedback, but still not very enthusiastic was T6 who commented that error feedback required a lot of revision, while content feedback also required some improvements with these words:

To be honest, I found myself needing to revise most of the AI's feedback... even when it provided correct error codes, it often missed other errors in the same sentence or gave misleading feedback... I may prefer using content feedback, but I do not think that I will use it much.

Collectively, these responses include serious concerns for error feedbacks' reliability while expresses somehow positive attitude toward AI-generated content feedback, provided small to moderate intervention was performed to refine or supplement it. The structured content feedback was found useful, yet teachers' responses suggest that this could be improved with awareness of existing gaps in content feedback.

As another sub-category within the positive aspects of refining AI feedback, *Appreciation of AI's guiding tone in error descriptions* emerged from the participants' insights. Under this sub-theme, T4 refers to the sub-code Positive outlook on the guiding tone of AI feedback. She acknowledged the guiding tone in error descriptions as a positive feature but expressed accuracy as a barrier for the efficiency of them:

I liked AI's guiding tone sometimes. Actually, error descriptions would really be helpful if feedback was accurate because sometimes students don't understand them, like amm... the comments, I mean. And at those times, we usually provide some oral feedback to help them. However, AI could save us from this burden sometimes, I mean, if there was some accurate comment, I can say.

Additionally, another sub-code is *Promotion of teacher self-reflection for comprehensive feedback*. Only one teacher noted an additional benefit, which is not directly related to error or content feedback, but rather to the process of refining AI feedback as a teacher and how it influenced her own feedback practices. T5 reflected that the refinement process itself helped improve the depth of her feedback: “With AI's help, I ended up giving more comprehensive feedback than I usually do.” This suggests that despite its limitations, AI-generated feedback may play a supportive role in enhancing teacher-generated feedback.

The results suggest teachers do not trust error feedback and expressed the need for substation intervention for using it, whereas for content feedback, there found it more useful than grammar feedback with some small refinements to its feedback.

4.4.3 Teachers' attitudes for future usability of AI-generated feedback

This section delves into teachers' perspectives on the usability of AI-generated feedback and recommendations for its integration into classrooms. The interviews of teachers contributed to understanding teachers' key concerns, and their resolutions for them.

A key finding from the interview responses is that all eight teachers expressed reluctance to rely on AI-generated feedback in its current form, particularly in relation to error-coded feedback. While several teachers (T2, T3, T5, T7, T8) indicated a willingness to use content-related feedback after making necessary refinements, others (T1, T4, T6) remained hesitant about using it (See Table 4.24).

Table 4.24

Teachers' Attitudes for Future Usability of AI feedback

Sub-themes	Total Mentions (n=8)	Teachers' Mentionings
Reluctance to use AI feedback (mainly for error feedback)	8	T1, T6, T4, T8, T3, T2, T7, T5
Need for AI training for contextual Understanding and tailoring it for classrooms	8	T1, T6, T4, T8, T3, T2, T7, T5
Need for university-level AI training support	2	T1, T5

One of the major sub-themes that emerged from the data was *reluctance to use AI feedback*, particularly for error code feedback. All eight teachers expressed reluctance to use AI for error feedback in its current form. T3 firmly suggested not using it in this current form and reasoned: "Writing isn't just about correct sentences. It's about clear, logical communication, I believe." This suggests T3 does not consider AI really take this into consideration, and its approach of feedback is grammar-focused rather than meaning-focused.

T8 also stated her preference for content feedback incorporating missing elements while expressing strong reluctance to use it for error feedback: "I would use it definitely, but

not for error feedback. Actually, I found it more useful for content feedback, though I'd say, I still add my own refinements for missing elements" In similar vein, T7 also suggested using content feedback is more likely for her because of ideas offered by AI are useful from her perspective, but she also conditioned that she has to add her own insight for content:

I would use it for, let's say, content feedback, and while adding my comments for content development and organization. Yes... good, yeah... it provides good things, but it needs teachers to look at it before giving feedback.

Similarly, T5 explained, "In terms of accuracy issues, I wouldn't use it for error feedback, but I would only use it for content feedback maybe." T2 also disapproved use of it for error feedback for now, and supported the use of AI for content feedback, indicating that she would selectively use it with her own additions:

maybe only for providing content feedback I can use it because accuracy was better than the error codes feedback, let's say. But it would be again with my refinements and additions. However, I definitely think it can help me in class. Umm... as I said, for error, I can say no, for now

Unlike other teachers, T4 expressed a strong disapproval for using AI feedback for both error codes and content feedback, considering the major verification process required:

To be honest, I think not now because I need to double-check everything. I needed to check whether the content feedback is correct or whether the error codes are correct. So, this seems a bit complicated for me right now... Content is better, I guess, but still not enough...

Similarly, T1 and T6 also stated disapproval for the use of it for content feedback. T1 stated, "I don't feel like using it... I really personally give importance to giving content, but when I looked at the feedback, most of the content feedback was either missing or overlooked." T6, similarly stated "not positive about it."

Hence, the subsequent subtheme all teachers commonly mentioned is that there is a strong need *Need for AI training for contextual Understanding and tailoring it for classrooms*. Teachers' reasons for this training had different focus points: the need for contextual understanding, grasping meaning and nuance, alignment with teaching methods, and institutional and technical support for teachers.

To start with, the need for training was echoed by T1, who asked during the interview, "There is a serious need for AI training, right?", and by T7, who similarly remarked, "...the system needs several improvements." Another possible approach reported by T4 is to train AI on how to read student papers with a focus on meaning: "Maybe AI needs more training on how to read student papers because it focuses too much on errors

sometimes and risks overlooking meaning.” T8 also listed specific areas in need of improvement, such as “better detection of meaning level, word choice, content, context, [and] academic tone.” As can be seen, these levels are also concerned with meaning and contextual understanding, which teachers consider AI fails to address oftentimes.

Furthermore, within the same theme, T6 also highlighted the need to tailor AI to classroom needs for better contextual understanding of student papers. T6 commented on his reason for university-level training as follows: “We need a system that can better understand context, our students' context, to provide more accurate error feedback.” Similarly, T3 emphasized the importance of aligning AI feedback with classroom teaching methods and advocated for training AI accordingly: “AI needs to analyze writing... the way we teach it.”

As for the form of training, two teachers suggested the *Need for university-level AI training support*. T1 suggested that professional, university-level support would be more effective than individual efforts: “...instead of individual efforts, the university can help us with this kind of AI training.” Likewise, T5 emphasized the need for institutional and technical support to improve and localize AI tools based on each institution's specific understanding of feedback:

If there was a unit at the university that could work on improving the accuracy of AI feedback, specifically to fit it to our teaching practices, or tailoring it, let's say using real examples from our school.

4.4.4 Teachers' improvement suggestions for AI-generated feedback

All these findings suggest that AI training needs improvements in its feedback approach to reach feedback similar to that of teachers. Teachers communicated several valuable suggestions to improve AI systems to better resemble teachers' feedback. The analysis of interviews with eight educators revealed consistent patterns in their recommendations, highlighting specific areas where AI feedback mechanisms could be enhanced. These suggestions focus primarily on shifting AI's emphasis from surface-level corrections to deeper content analysis, improving contextual understanding, better identifying repetitive elements, and providing more actionable guidance (See Table 4.25).

Table 4.25*Teachers' Improvement Suggestions for AI-generated Feedback*

Sub-themes	Total Mentions (n=8)	Teachers' Mentionings
Prioritize meaning over grammar	7	T1, T6, T4, T8, T3, T2, T7, T5
Better detection of context problems and solutions	4	T1, T6, T8, T5
Better detection of repetitive content and solutions	4	T8, T3, T2, T7
Better detection of missing feedback and solution	1	T6
Solutions for providing specific/ actionable ideas for content	1	T2

In the category of improvement suggestions, *Prioritize meaning over grammar* appeared as one of the major sub-themes. Seven out of eight teachers suggested that AI should prioritize meaning over grammar in its feedback approach. T3 stated, "AI needs to go beyond grammar and assess effective communication... Fixing grammar won't help if the sentence doesn't make sense." T2 also recommended a meaning-first approach in feedback provision: "AI should check meaning and context first before grammar."

T5 similarly advised the same approach, suggesting that it is the one followed by teachers in classrooms: "Like we do, AI should check meaning first, not grammar because as we do in school, we always give feedback for meaning first and grammar later." T7 also suggested a shift from grammar-focused to meaning-focused feedback: "The most important improvement should be within meaning, context, and logical sentence flow over grammar."

Six teachers emphasized *the need for better detection of contextual problems and solutions* as another sub-theme. T8 suggested, "It should not see a sentence in isolation but always consider each sentence in relation to the one before and after it." This suggests that AI's reading approach can be problematic and needs to be improved. T5 also proposed aligning AI feedback with the instructional writing goals of institutions to avoid unnecessary suggestions: "Given our student profile and context, I think it is unlikely that students have the background knowledge to add statistics because that would require prior research."

The sub-theme of *Better detection of repetitive content and solutions* also emerged as another important finding. To improve AI's ability to provide feedback for repetitive content, teachers also shared several suggestions. T2 stated, "AI should compare body paragraphs to check for repetitive content." As another solution, T7 recommended, "It should not overlook errors related to meaning and should consider context and the logical flow between sentences. This way, it could identify whether ideas are too similar, repeated, or unrelated." T3 suggested that AI should detect "circular reasoning", for example, track repetition across paragraphs, and detect underdeveloped arguments." He stressed the importance of higher-order level feedback as a key nuance provided by teachers and argued that AI should be rendered capable of doing the same to achieve teacher-like feedback. As a revision strategy, T2 suggested, "If [AI] detects similar ideas, it should give students the option to either write something different or combine the paragraphs." This could provide students with the chance to consider their choices and decide accordingly.

As for the sub-theme of *better detection of missing feedback and solution*, T6 stated, "AI needs to... provide feedback on all aspects of the essay, including the conclusion... If the word limit is a concern for AI, it should not be restricted." The word limit for AI was confined to 200 words for the purpose of the study, but the teacher suggested that this limit might have been the reason AI failed to provide feedback on all aspects of the essay. This suggests that not limiting AI in content feedback might result in better and more complete AI-generated feedback.

The last sub-theme, entitled *Solutions for Providing Specific/Actionable Ideas for Content*, emerged from the analysis. As a suggestion to address this issue, T5 shared an approach she also implements herself: "Usually I quote their sentences and ask them to explain. Or if the idea needs expansion, I add a note and say, like, 'You need to add ideas, give examples' kind of thing for this idea." This approach is widely adopted in classrooms, yet with the current capabilities of AI, offering such personalized suggestions is not very feasible at the moment.

Taken together, these findings suggest that AI feedback systems require further development to emulate teacher-like reasoning, particularly in areas such as prioritization of meaning, detection of contextual issues, identification of repetitive content, and the provision of more specific and actionable suggestions for content development.

5 CONCLUSION, DISCUSSION, IMPLICATIONS AND SUGGESTIONS FOR FURTHER STUDIES

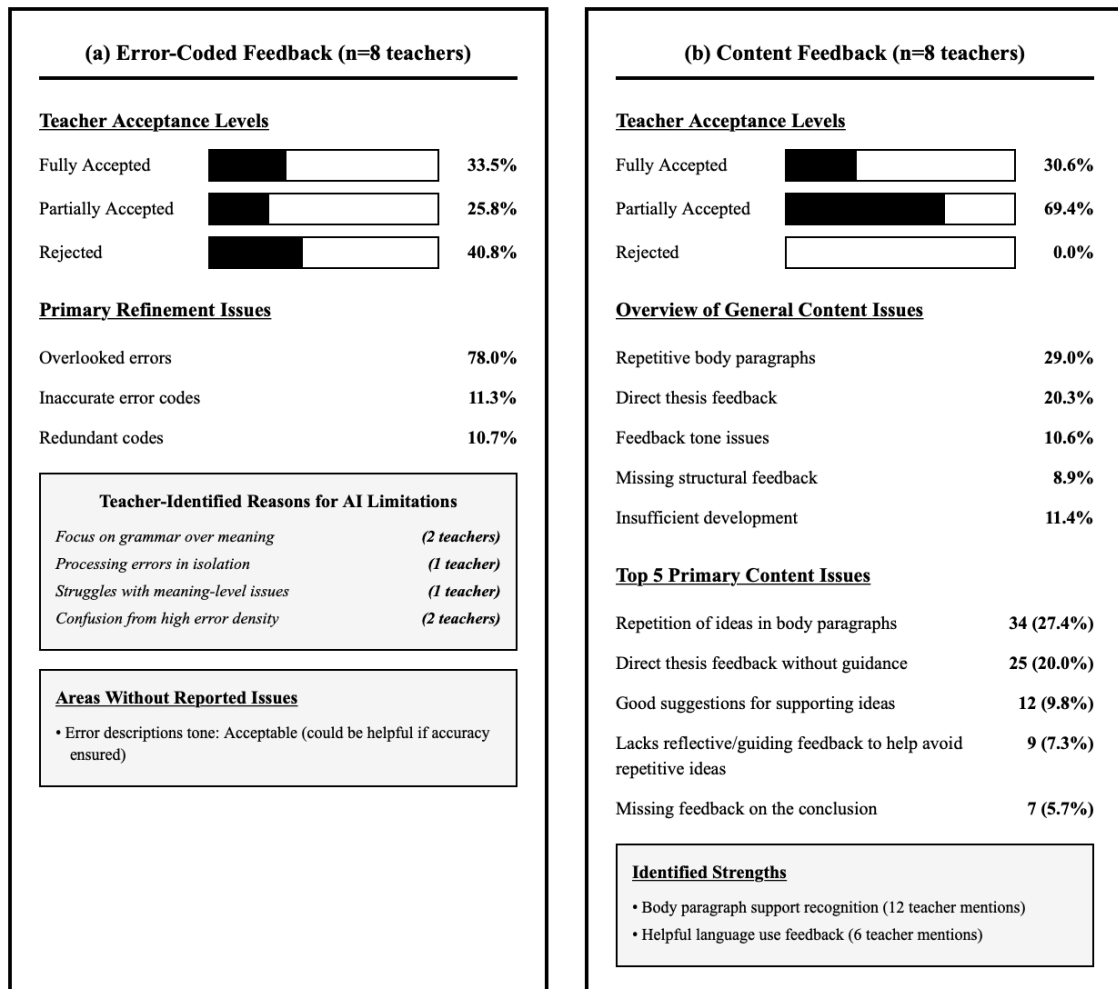
5.1 Conclusion

This study was designed as a qualitative, holistic multiple case study aimed at uncovering how preparatory class teachers refine AI-generated feedback for student essays on both error and content dimensions. It also explores their perceptions and evaluations of the teacher-like accuracy of such feedback, their insights into the advantages and challenges of refining it for classroom usability and their views and suggestions for its effective classroom use and further improvement. To address these goals, data were collected through teachers' color-coded reviews (including teachers' edits), structured reviews (documentation of their acceptance levels of AI feedback, rationales, teachers' notes of edits), accuracy evaluations, and individual interviews to ensure triangulation of data from various sources and gain deeper insight into the research phenomenon. The data was analyzed using the content analysis method by Hsieh and Shannon (2005), combining both summative and conventional approaches depending on the nature of the data. In this vein, summative content analysis was applied to data derived from pre-defined categories of color-coded reviews of teachers, and the review table entries such as acceptance levels of AI-generated feedback and teacher rationales for refining feedback. On the other hand, conventional content analysis was used for open-ended responses, including teacher notes of edits for both error-coded feedback and content feedback, and interview transcripts, where codes and themes were generated inductively from the raw data. For the rubric data, consensus ratings were used to interpret teachers' overall evaluation of AI feedback, and rubric commentaries were used to contextualize teachers' justifications.

To provide a clearer overview and understanding of all findings, the following diagram presents the count and percentage of findings categorized under emerging themes. These findings resulted from the integration of multiple data sources throughout the study, which complemented each other both quantitatively and qualitatively (See Figure 5.1).

Figure 5.1

Summary of Findings: Accuracy, Limitations, and Strengths of AI-Generated Feedback as Perceived by Teachers



Note. Data represents teacher evaluations of AI-generated feedback across two dimensions: error-coded feedback and content feedback. Acceptance levels indicate the degree to which teachers found the feedback suitable for direct classroom use without modification. Issue percentages represent the frequency of specific problem identified during teacher refinement processes. Strengths represent positive aspects of AI feedback recognized by teachers during evaluation and refinement activities.

According to the findings illustrated above in figure 5.1. above, prep-class teachers' evaluations and refinement of AI-generated feedback revealed significant limitations in AI's ability to emulate the teacher-like feedback desired for a possible classroom use. The main limitations were perceived in the comprehensiveness of error detection and prioritization of some issues, that is, in its address of student errors and aspects of content writing. While some promising aspects were identified, teachers believed that AI feedback needs improvement before it is shared with students.

Regarding teacher acceptance levels of AI-generated feedback shown in Figure 5.1, the findings indicated that teachers were mostly critical of the accuracy of AI-generated feedback, especially in the case of error-coded feedback. Specifically, only 33.5% of error feedback was fully accepted by teachers, while 25.8% was partially accepted and 40.8% was rejected. This shows that most teachers identified major flaws during their refinement processes, as seen in both their color-coded and structured reviews. This need for substantial revision was also highlighted in their rubric comments and interviews, where teachers frequently referred to overlooked RW codes and other major issues that made it difficult to fully correct problematic parts

The findings also indicated that accuracy was not affected by whether the identified error codes were correct or not, but was seriously impacted by overlooked errors that went unnoticed and received no feedback. This suggests possible processing issues, particularly in cases with a high density of errors. Teachers observed during their reviews that student papers in real classroom contexts can be just as densely erroneous as the papers used in the study, and that AI is not yet capable of processing such texts effectively.

Teachers also identified several possible underlying reasons for the limitations in error-coded feedback. Based on interview data, these included the AI's excessive focus on grammar over meaning (mentioned by two teachers), focusing on errors one by one in isolation (one teacher), failure to detect meaning-level issues (one teacher), and confusion caused by high error density (two teachers). These issues contributed to the high rate of overlooked errors (78% of all refinements), inaccurate codes (11.3%), and redundant codes (10.7%).

As for content feedback, the results showed that teachers' perceptions were more positive than they were for error-coded feedback, as the content feedback generally did not contain serious accuracy or relevance issues. In fact, none of the content feedback instances were rejected. However, full acceptance was also limited to 30.6%, while more than half (69.4%) of the content feedback was only partially accepted. This suggests that although content feedback was seen as more accurate and usable by teachers, it still required some improvement. Most notably, the absence of feedback on certain key points or components of the essay, which teachers felt should have been prioritized, often led to partial acceptance.

Importantly, none of the teachers rejected the content feedback entirely, suggesting that they found it usable in class and beneficial for students. However, the high rate of partial acceptance (69.4%) shows that improvements are still needed. From the teachers' perspective, the most prominent gap in content feedback was the failure to address repetitive ideas in body paragraphs, which often featured similar or identical content without variation in examples, content, or topic sentences. This issue accounted for 29.3% of all problems identified in content feedback. Other notable gaps included direct thesis feedback without guidance (20.3%), tone-related issues (10.6%), missing structural feedback (8.9%), and insufficient feedback for body paragraph development (11.4%).

As also illustrated in Figure 5.1, the top three issues mentioned by teachers concerning content feedback were: AI's failure to detect repetition in body paragraphs (27.4%), the use of direct thesis feedback without guidance (20.0%), and the lack of reflective or guiding comments to help students revise repetitive content (7.3%). Additionally, missing feedback on conclusions (5.7%) indicated that the AI system sometimes overlooked key aspects of essay structure, organization, and content development.

On a positive note, several teachers acknowledged that the content feedback contained some useful ideas. Specifically, they highlighted that the AI provided helpful suggestions for supporting ideas (9.8%), which they believed could assist students in improving underdeveloped points in their writing. These strengths contributed to greater teacher openness to using content feedback compared to error-coded feedback. This is a noteworthy finding, as teachers reported no positive aspects for error-coded feedback, whereas they did for content feedback. In the content analysis of teacher review notes, these strengths, particularly the recognition of body paragraph support points (mentioned twelve times), indicated that teachers found it relatively useful. Additionally, teachers made positive comments six times regarding AI's language-related feedback, noting its usefulness in improving word choice, clarity, or academic tone.

As for refining AI-generated feedback overall, teachers reported several concerns that made them reluctant to use it. To provide a clearer overview of these challenges and the positive aspects of the refinement process, the following Figure 5.2 presents a summary of findings drawn from interview data.

Figure 5.2

Summary of the Pros and Cons of Refining AI-Generated Feedback, Usability of Feedback Based on Teachers' Views, and Their Suggestions for Improvements

Pros & Cons of Refining AI Feedback (Interview Data)	Usability of AI Feedback & Teacher Attitudes (Interview Data)
Challenges & Concerns Need for substantial teacher refinement (all teachers) <i>Increased time investment for refining AI feedback</i> (4 teachers) <i>Increased cognitive load due to interpretation of AI feedback</i> (3 teachers) <i>AI feedback's lack of human judgment and nuanced understanding</i> (5 teachers) Positive Aspects of Refining AI Feedback <i>Easier refinement of content feedback compared to error feedback</i> (5 teachers) <i>Appreciation of AI's guiding tone in error descriptions</i> (1 teacher) <i>Promotion of teacher self-reflection for comprehensive feedback</i> (1 teacher)	Current Teacher Attitudes Strong reluctance to use AI feedback (mainly for error feedback) (all 8 teachers) Greater openness to using AI-generated content feedback with refinements (5 teachers) Prerequisites for Future Use Need for AI training for contextual understanding and classroom tailoring (all 8 teachers) Need for university-level AI training support (2 teachers) Teacher Suggestions for Improvements Prioritize meaning over grammar (7 of 8 teachers) Better detection of context problems and solutions (4 teachers) Better detection of repetitive content and solutions (4 teachers) Better detection of missing feedback and solutions (1 teacher) Solutions for providing specific/actionable ideas for content (1 teacher)

Note. Table summarizes interview responses of teachers and summarizes their views on refinement challenges, usability, and improvement suggestions.

The findings of the interview data revealed that all eight teachers reported that refining AI-generated feedback was nearly as time-consuming and cognitively demanding as providing feedback from scratch, particularly for error-coded feedback. This shared concern was primarily due to the substantial revisions required for such feedback. Specifically, four teachers mentioned increased time investment, three referred to the heavier cognitive load caused by interpreting AI feedback, and five highlighted the lack of human judgment and nuanced understanding in AI-generated responses.

Despite these challenges related to the refinement of error-coded feedback, the refinement process for content feedback was viewed more positively. A majority of teachers (five out of eight) evaluated content feedback favorably and reported that it was easier to refine. Additionally, several teachers appreciated the refinement process itself, as it encouraged them to reflect on their own feedback practices, ultimately contributing to the generation of more comprehensive feedback.

Regarding their future willingness to use AI-generated feedback and integrate it into their classrooms, teachers were generally not favorable, particularly due to the extensive intervention required for error-coded feedback. In this regard, teachers appeared to fall into two distinct groups: those who were reluctant to use AI-generated error-coded feedback in its current form, and those who were more open to selective use. The latter group expressed more positivity toward content feedback, provided that only minor adjustments were needed. In short, there was a strong and unanimous reluctance among all eight teachers to use error-coded feedback, while five teachers showed greater openness to using content feedback after some refinements.

As for the necessary conditions for classroom use, teachers proposed two key improvements: first, training AI to develop contextual understanding of classroom feedback and student needs; and second, offering professional development support at the university level that takes into account current feedback practices and student profiles. All eight teachers emphasized the need for AI to be trained specifically for classroom contexts, as current systems lack sufficient contextual awareness to provide meaningful and pedagogically sound feedback. Two teachers, in particular, called for institutional support for training AI tools to better align with university-level educational settings. These steps were seen as necessary for fine-tuning AI feedback and possibly adopting it for specific tasks within the limits of its current capabilities and algorithms.

Regarding their broader suggestions and expectations for future usability and integration of AI-generated feedback in classrooms, teachers shared several implications based on their refinement experiences and observations of AI limitations. For error-coded feedback, seven out of eight teachers emphasized the need for AI to prioritize meaning over grammar. For content feedback, teachers expressed expectations such as better detection of context-related issues (mentioned by four teachers), improved recognition of repetitive content and appropriate suggestions (four teachers), better identification of missing feedback (one teacher), and clearer, more actionable suggestions (one teacher). In summary, teachers emphasized that for AI-generated feedback to be effective, it must prioritize meaning, interpret sentences holistically rather than in isolation, and for content feedback, assess writing at the paragraph and essay level rather than focusing solely on individual sentences.

5.2 Discussion

This section aims to synthesize previous literature findings with the outcomes of the present study. First, it focuses on teachers' refinement patterns, needs, and pedagogical insights derived from the research findings and discusses how they corroborate with previous research. Second, findings regarding teachers' evaluations of AI feedback accuracy are explored and compared with existing literature. Finally, the challenges and advantages of refining AI feedback due to various factors are discussed. The findings derived from this part are also interpreted in the light of relevant literature.

5.2.1 Teachers' Refinement of AI Feedback: Patterns and pedagogical insights

The first research question investigated how preparatory class teachers refined AI-generated feedback using their own expertise and insights. A key finding of the study was the significant level of teacher intervention required to make AI-generated feedback suitable for student use. This outcome was primarily caused by the high number of overlooked errors in the AI feedback, particularly in error-coded feedback, which accounted for 78% of all refinements. This finding points to a critical issue in AI's error detection process. Specifically, the limitation does not lie in the accuracy of error identification when errors are detected, but in the comprehensiveness of its detection across student writing. The types of refinements implemented by teachers illustrate why some errors were overlooked.

While previous studies did not specify the types of errors that AI addressed or those it failed to detect, the present study enabled the identification of specific error types that were overlooked by AI. The most notable finding was the consistent omission of RW (rewrite) code instances. Teachers classified this overlooked error as global-level error because it directly affect the sentence meaning. In this study, it was seen rewrite code was not assigned despite students' sentence contain multiple simultaneous issues, sentence fragmentation, unclear meaning, or poor structure.

The primary issue noted by teachers was that AI tended to assign individual error codes that were insufficient for addressing global issues requiring the RW (rewrite) code, as they failed to resolve the underlying meaning-related problems. This suggests that AI systems prioritize grammatical correctness over meaning and clarity, which contradicts Ellis's (2009) principle that global errors, that is those affecting meaning should take precedence over local, surface-level ones. This grammar-focused orientation also stands

in contrast to contemporary writing instruction philosophies, which emphasize communication over form. In this respect, the present findings also concurred with Searle's (1980) Chinese Room Argument, which asserts that AI manipulates symbols without true understanding. Since the errors in this study were symbolically coded, this theoretical perspective helps explain why AI in the present study struggled significantly with meaning-level and context-dependent issues: it does not truly understand the context of communication.

Another notable finding was the frequency of incorrect or inconsistent code application (11.3%), particularly in cases involving word choice and word form distinctions. Teachers reported that the AI often flagged word choice issues using inaccurate coding. This finding was also echoed by Algaraady and Mahyoob (2023), who stated that AI's mislabeling arises from its inability to understand the meaning, structure, and pragmatics of writing. This suggests that an excessive focus on grammatical correctness leads to the miscoding of word choice problems, for instance; labeling them with the word form code while disregarding the underlying meaning issues.

The third category of refinement involved redundant codes (10.7%), where teachers removed unnecessary feedback. In their findings, Pang et al. (2024) also referred to this issue, stating that unnecessary or irrelevant information was often produced by the AI. This problem also stemmed from incorrect interpretation or reading of the sentence. This insufficiency in semantic understanding became apparent when teachers added their own codes during refinement and noticed that the AI's feedback had been misplaced or unnecessarily applied. While quantitatively less frequent, refinements related to redundant codes further support the argument that AI struggles to understand the meaning of a sentence.

In terms of content feedback, the findings of this study can be considered relatively favorable compared to error-coded feedback. While the majority of teachers still identified areas needing refinement, the overall perception was more positive. However, the results are not as strongly favorable as those reported by Dai et al. (2023), who suggested that AI-generated feedback closely resembles teacher "textual content" feedback in aspects such as goal, topic, novelty, and clarity. In their study, *topic* referred to the relevance and development of ideas, including appropriate elaboration, while *novelty* referred to the originality and variety of ideas, avoiding repetition.

In the present study, however, repetition of ideas in body paragraphs emerged as the most frequently mentioned issue, raised in 34 teacher comments (27.4%). This suggests that AI fails to detect problems in idea development when it overlooks repetitive content. Thus, this finding contradicts Dai et al.'s (2023) conclusion regarding AI's ability to provide teacher-like feedback on topic development.

On the positive side, the present study also highlighted strengths in AI-generated content feedback. For instance, teachers in 12 cases (9.8%) noted that the AI provided useful suggestions for supporting ideas, particularly in improving body paragraph development. This finding indicates some level of agreement with Dai et al.'s (2023) study, showing that AI can sometimes offer constructive suggestions related to content elaboration.

However, the findings of the present study closely corroborate Yoon et al.'s (2023) discovery regarding GPT-3.5, which showed that AI often overlooks major content issues such as repetitive ideas at the paragraph level while focusing instead on word-level repetition. Their study further reported that GPT failed to detect coherence and cohesion issues, especially when no problems were perceived in terms of grammatical correctness. This finding is consistent with the results of the present study, in that the majority of teachers expressed that they had to add missing feedback related to content coherence and organization.

Teachers observed that Claude 3.5 Sonnet prioritized offering supportive ideas and language points, which was generally perceived as a positive feature. However, they emphasized that the model should instead prioritize addressing more critical gaps in student writing that affect organizational coherence, such as repetitive structures (e.g., identical ideas across body paragraphs) and missing components, including conclusions that lack a restatement of the main idea or a final thought. In their perspective, missing elements in structure of the essay or poor coherence should take precedence over the use of supporting ideas. This is because, even when strong supporting points are present, overall communication may fail without clear structure and coherence.

This observation was in line with the study of Ellis et al.'s (2008) description of effective feedback, which states that feedback should be selective and help students prioritize key issues. In the present study, it was observed that teachers consistently emphasized the need for more nuanced feedback that selectively targets the most important content-related problems.

Most of the teachers also expressed their concerns about direct thesis feedback without guidance, reported in twenty-five instances (20.03%). Although it was not typically referring to direct thesis statement provision, Obaidoon and Wei (2024) also discussed that AI has issues with corrective guidance, especially when dealing with higher-order content feedback elements. This study also identified actionable guidance as one of the shortcomings of Claude, and the present study aligns with this finding. Several teachers mentioned the lack of reflective or guiding feedback to help students avoid idea repetition (nine mentions; 7.3%) and missing feedback on the conclusion paragraph (seven mentions; 5.7%). These findings corroborate the results of Obaidoon and Wei (2024), who reported that Claude failed to provide metacognitive prompts that would help students revise their ideas more independently.

These findings can be interpreted in three main ways. First, AI systems are not ready to manage student papers with a large proportion of meaning-level errors because their approach to error-coded feedback is excessively form-focused. This issue was also felt in the content feedback, where teachers reported that real problems in student writing were not prioritized due to the system's lack of holistic, paragraph-level processing. This became especially clear to teachers when they noticed during their reviews that identical ideas across body paragraphs were not recognized.

Second, understanding what students actually mean in their sentences presents an additional challenge, particularly in the context of this study, where student writing often included Turkish language transfer issues as well as developmental errors common among B1-level learners. These learners frequently carry fossilized errors from earlier stages of language acquisition. Such cases require the kind of nuanced understanding that experienced teachers bring, especially those familiar with these error types. Similarly, this level of nuanced teacher reasoning was also absent from the AI-generated content feedback, which did not provide detailed support for content development.

Lastly, limitations in the processing capacity of AI, due to current technological constraints and algorithmic design, may also have contributed to these shortcomings. To conclude, the findings of the present study concur with both theoretical and empirical research, despite some subtle differences in feedback tasks and study contexts.

5.2.2 Teachers' evaluations of the accuracy of AI-generated feedback

In line with the teacher refinement results, teachers' evaluations also indicated a significant gap between current AI capabilities and teacher expectations, with evaluations of low acceptance rate of AI error feedback (33.5% fully accepted, 40.8% rejected) for error-coded feedback generated by AI. Accuracy issues were mainly caused by major overlooked errors including not detecting rewrite code needing sentences.

The higher acceptance rate for content feedback (30.6% fully accepted; 69.4% partially accepted) suggests that AI may be more effective in providing general content guidance than in delivering specific error correction. This finding resonates with Kurt and Kurt's (2024) study, in which both GPT and teacher feedback were preferred over peer feedback. A similar preference for content feedback was also expressed by the teachers in the present study.

However, in contrast to the same study's findings mentioned above, teachers in this study indicated that while AI-generated content feedback may identify certain strengths or issues, it often fails to address major organization-level problems and lacks the guidance necessary for effective formative feedback. This shortcoming directly contrasts with Nicol's (2010) principles for effective feedback, which emphasized the notion that feedback should be selective, specific, and contextualized. Similarly, in their mixed method study comparing teacher and AI feedback of student writing in terms of quality and quantity, Lu et al. (2024) concluded that GPT's suggestions are less clear and specific compared to those of teachers. Teacher comments were found to be more detailed and specific, covering different areas (e.g., grammar, organization, coherence). In the present study, teachers also raised the same concerns regarding AI-generated content feedback, arguing that it sometimes lacked nuanced, personalized, and targeted suggestions essential for the development and completion of ideas in student papers.

The teachers' rubric commentaries provided valuable insight into why AI-generated feedback received varying accuracy ratings. The unanimous high rating for Paper 6 particularly highlights this point through teachers' explanations. For example, T7's observation that the AI places "too much focus on grammar" and "couldn't detect that the sentence has meaning problems" further supports Chomsky's (2023) recent theoretical position, published in *New York Times Magazine* regarding AI's fundamental lack of true understanding of language principles.

Similarly, in the interview data, T4's insight that AI focuses on "errors one by one in isolation," rather than considering their interrelationships, helps explain why AI often misses meaning-level issues. This also aligns with Guo and Wang's (2023) finding that AI-generated feedback lacks contextual awareness of how different elements of writing work together.

Another issue affecting the teacher-like accuracy of AI feedback is its tone, which sometimes proposes exact solutions in content feedback. This clearly goes against the nature of metalinguistic feedback. Theoretically, this can be explained by Vygotsky's (1978) concept of scaffolding, which emphasizes that effective learning occurs when students receive guided support within their zone of proximal development—helping them gradually become independent. Reflecting this, T6 commented on AI's direct suggestions for thesis statements, stating that such feedback "will not help students think and fix their own mistakes." However, apart from this instance, T6 perceived most of the AI-generated content feedback as guiding rather than overly directive, suggesting that the AI has the potential to provide support in a more scaffolded, teacher-like manner.

As for the error descriptions in error-coded feedback, some teachers appreciated the AI's guiding tone. This is a noteworthy and positive finding, especially given that several recent studies have raised concerns about the tone of AI-generated feedback. For example, AlGhamdi (2024) reported that students found GPT's tone to be overly formal or emotionally distant. Similarly, Kurt and Kurt (2024) noted a lack of human warmth and interpersonal tone in AI-generated feedback. However, in the present study, teachers did not raise such concerns. On the contrary, some even valued the AI's supportive tone in error descriptions, emphasizing its potential to reduce teacher workload when the feedback provided by AI was accurate.

5.2.3 Challenges and advantages of refining AI-generated feedback

The difficulties teachers experienced in this study while trying to improve AI-generated feedback raise important questions about the practicality of using this technology in everyday teaching. The findings revealed that refining AI feedback required considerable time investment and cognitive effort, particularly with error-coded feedback. Teachers often had to check for meaning-level issues and assign rewrite codes, which increased their workload rather than alleviating it. However, they expressed a relatively more positive stance toward content-related feedback.

Although content feedback lacked some important nuances, teachers found it easier to refine and more pedagogically useful as a starting point for developing more effective comments. In the present study, teachers mainly refined AI content feedback to address repetitive or identical content across body paragraphs or missing elements in the essay's organizational structure. This finding aligns with Farrokhnia et al.'s (2023) SWOT analysis of GPT-3, which noted that GPT could provide broad, generic comments but struggled to offer truly meaningful suggestions due to limited contextual understanding. However, the present study partially diverges from this, as teachers found the comments relevant and meaningful for the development of student ideas. While they did not suggest removing any of the comments, they emphasized that prioritizing and elaborating on key points could enhance effectiveness of feedback provided.

Teachers also noted that the feedback occasionally felt mechanical or template-driven, not because of its tone, but because it failed to prioritize the real needs of the student paper. Nevertheless, AI feedback in this study was not criticized for clarity or tone, which had been issues with other platforms in the same research context. This suggests that while the AI feedback had pedagogical value, human input remained essential for fine-tuning nuances and supporting deeper student engagement in revision.

Overall, recent advancements in AI feedback systems have made content feedback more usable and supportive, especially for developing ideas. Yet, teacher guidance remains essential to maximize pedagogical potential and address issues such as repetitive content or organizational gaps. This aligns with Wilson and Czik's (2016) findings that combining AI and teacher feedback enables instructors to focus more on higher-order writing skills. Some teachers in this study echoed this, appreciating the structured format of AI content feedback, particularly its guiding questions that prompted student reflection and supported idea development.

Different from the rest of teachers, T5 only saw positivity in refining AI-generated feedback, noting that refining AI feedback can lead to more comprehensive feedback than teachers might typically provide. This suggests that AI can serve as a helpful scaffold and offer a structured baseline that helps teachers notice areas they might otherwise overlook. Guo and Wang (2024) also found that GPT produced richer feedback when paired with teacher input.

However, these findings must be interpreted cautiously. Teachers in this study were asked to filter and improve AI feedback, which may have heightened their sense of responsibility and led to a tendency to overcorrect in order to maintain their pedagogical authority. This concern was also raised by Zawacki-Richter et al. (2019), who warned that teachers may act in tendency to emphasize their pedagogical expertise.

Moreover, the research context required teachers to invest significant time and attention in reviewing AI feedback, something that is often unfeasible in typical classroom conditions. As Carless (2006) noted, teacher feedback in real classroom settings is often limited in depth and clarity due to time constraints and institutional pressures. In contrast, teachers in this study engaged in more detailed reviews than they typically would, meaning that the number of overlooked errors identified here may be exaggerated relative to normal practice.

The study also revealed that the refinement process could create confusion and add to the cognitive and emotional burden. Yu et al. (2021) emphasized that providing feedback is emotionally demanding and often influenced by teachers' emotional states. While AI-generated feedback is free from human bias, it is not free from accuracy issues, and its inability to prioritize effectively can lead to discrepancies between teacher evaluations and AI comments. These gaps can increase uncertainty and frustration among teachers during refinement process, who still feel responsible for ensuring the feedback is accurate and actionable for students. Yet, it is important to recognize that such divergences are very natural and also occur between teachers. As Sadler (1989) pointed out, teachers often differ in their ratings and comments on the same student paper, which is why calibration sessions are necessary to standardize evaluations.

In conclusion, the process of reviewing and refining AI-generated feedback, especially error-coded feedback, was described by teachers as time-consuming and cognitively demanding due to presence of considerable number of overlooked errors and interpretation challenges. This did not address long-standing shortcomings of teacher feedback highlighted by Brown (2007), such as inconsistency and delays caused by fatigue, mood, and subjectivity. Instead, the study showed that refining AI feedback adds an extra layer of responsibility for teachers and may intensify existing pressures. As Dimitriadis and Gašević (2021) noted, large class sizes already challenge the timely and dialogic delivery of feedback. Although AI-generated feedback is scalable in theory, the

limitations in accuracy and interpretability of error codes, as well as prioritization issues in content feedback, did not reduce teachers' overall time and cognitive burden in this study.

5.2.4 Future classroom usability of AI-generated feedback

Another insightful finding from the interviews was teachers' specific recommendations for enhancing AI feedback's classroom usability which provides valuable insights for AI literacy development as well. While teachers expressed significant concerns about AI's classroom-readiness, they also offered concrete suggestions for improvement that can help foster teacher AI literacy on the levels of knowing and understanding AI, applying AI applications, evaluating AI applications, and AI ethics (Ng et al., 2021; Zhao et al., 2022).

In terms of understanding and evaluating AI, teachers pointed out that AI-generated feedback often failed to prioritize meaning and lacked paragraph-level contextual awareness. Based on their reviews, their recommended recalibrating AI feedback to prioritize meaning over grammatical issues, developing more sophisticated contextual understanding capabilities, improving detection of repetitive content, and providing more comprehensive feedback on structural organization and missing essay components. It can be said that this recommendation offered by teachers can be linked with Vygotsky's (1978) social constructivist theory, which emphasizes the importance of social context in learning, and assistance of a more knowledgeable other. And yet the claims that the "More Knowledgeable Other" (MKO) role can be attributed to AI, as proposed in Steiner and Mahn's (1996) study, seems implausible for the time being. Current AI systems fall short of this ideal, as they demonstrated poor performance in terms of contextual and meaning-level awareness. One of the commonly cited shortcomings of teacher feedback is that students often revise their drafts without fully understanding the mistakes identified. Unfortunately, due to the high number of overlooked error codes, AI-generated feedback did not effectively compensate for this gap as well.

Teachers also raised ethical concerns about the responsibility involved in AI refinement. Teachers highlighted the ethical concern that AI improvement should not rely solely on individual instructors and called for institutional-level collaboration and responsibility. They recommended systematic and context-aware AI training at the institutional level because placing the entire burden on teachers for training of AI can be unmanageable for

them and may lead to effective outcomes. The need for teacher training has also been emphasized in previous studies. For example, Scott-Branch et al. (2023) advocated for structured, context-aware AI training at the institutional level. O'Dea and O'Dea (2023), highlighted the the urgency of equipping graduates with AI-related skills by teaching them to critically assess AI-generated content and engage in hands-on practices to develop effective prompts. Teachers in the present study built on these findings, stating that teacher training efforts should be collaborative, involving both individual educators and institutional support for better outcomes.

Based on the findings, it can be concluded that this collaboration would be more beneficial if it focuses on collecting and analyzing teacher refinements in a way that contributes to the long-term development of AI systems.. This insight suggests the need to accumulate a representative dataset over the long term and implement a stepwise and strategic process involving both teachers and institutions. A similar approach is exemplified in Ninaus and Sailer's (2022) framework, which outlines a model for improving AI feedback through three essential steps: data recording, pattern detection, and adaptivity. Interestingly, teachers in the present study described a similar process when discussing how AI systems should learn from teacher refinements to enhance the quality of future feedback.

In line with the ethics and application components of AI literacy, teachers suggested using AI in a supportive role, particularly for content-level guidance. Whereas teachers unanimously showed reluctance and even resistance to using AI feedback for error-coded feedback owing to demanding effort of refining it. While using for content feedback, they also suggested using it for providing generic feedback rather than relying on it as a fully autonomous feedback provider. In addition, teachers emphasized the need for institutional-level training and collaboration, rather than placing the burden on individual instructors to train AI and integrate it into their classroom. They suggested that representative data should be systematically collected to train AI systems that can respond to the specific needs of the institution and its student profiles.

5.3 Implications and Suggestions for Future Studies

The findings of this study culminate in several implications and suggestions for future research. These findings can also guide educators, scholars, and AI developers in the integration of AI-generated feedback into classrooms, particularly in university

preparatory programs. The implications of this study can address multiple areas, including pedagogy, design, policy, future research directions for researchers. Moreover, based on these implications drawn from this research, several practical guidelines can be drawn for teachers and practitioners aiming to train AI to provide the type of feedback implemented in this study.

5.3.1 Pedagogical implications

As a general-purpose tool, AI's deployment as an educational tool is inherently limited, and so is its pedagogical expertise. Gen-AI differs from specialized AI tools such as Grammarly (Grammarly, n.d.) and Cambridge Write & Improve (Write & Improve, n.d.), which are specifically designed to provide feedback on writing. Notably, Cambridge Write & Improve goes beyond basic proofreading and differs significantly from Grammarly in its approach to error detection (Grammarly, n.d.). Unlike Grammarly, it incorporates a feedback loop and provides iterative comments aligned with CEFR levels. A clear understanding of these tools and their distinct features helps explain why current general-purpose AI tools may fall short of delivering the level of comprehensive error detection required in this study.

MacLeod (2019) discussed the features of Cambridge Write & Improve in detail, noting that unlike Grammarly, this tool does not offer direct corrections. Instead, it provides hints about potential errors, encouraging learner reflection and self-correction. The tool delivers feedback in up to four rounds, aligning feedback with CEFR levels and informing students of their progress. Besides, the system refrains from giving feedback when it is less than 90% certain of its accuracy, acknowledging that inaccurate feedback can be more detrimental in the long run. When uncertain, it highlights the entire sentence without explanation, signaling to the student that something needs attention. Moreover, by offering feedback in multiple rounds and initially focusing only on frequent errors, it avoids demotivating learners. This staged approach allows students to use the tool independently, addressing errors progressively and thus processing student writing more accurately.

The treatment of errors by these AI models highlights why Claude 3.5 Sonnet in the present study might have fallen short of detecting all errors comprehensively in a single cycle. The findings indicated that AI-generated metalinguistic (error-coded) feedback often required substantial teacher intervention, likely due to the high density of language-

related and meaning-level issues in student writing. This limitation in addressing all the things at once may be attributed to constraints in AI's technical or cognitive processing. In other words, unlike *Write & Improve*, which stages its error detection, starting from basic spelling and progressing to sentence structure and meaning-level issues, Claude 3.5 Sonnet in the present study attempted to handle all errors simultaneously. As a result, it may have overlooked deeper, context-dependent problems and instead provided piecemeal corrections at the word level. This could explain why a considerable number of errors went undetected. Perhaps, like *Write & Improve*, generative AI models could be trained to deliver feedback in stages, focusing on specific areas progressively to enhance their overall effectiveness.

A recent study by Corbin, Tai, and Flenady (2025) also explored why Gen-AI may not provide teacher-like feedback. The study questioned the role of Gen-AI as a feedback provider and proposed a recognition-based framework to evaluate and separate AI and teacher feedback in terms of quality. The framework classifies feedback into two types: recognitive and extra-recognitive. Based on this distinction, the study positioned AI feedback as extra-recognitive, limiting its capacity to general feedback that lacks pedagogical and contextual depth. It also lacks the ability to personalize feedback according to students' academic profiles. Such meaningful interactions can only be sustained by teachers who provide recognitive feedback, tailoring their comments to individual students' needs.

However, the same study also suggested that the pedagogical opportunity of Gen-AI feedback lies in its limitations. It can assist teachers by saving time on surface-level error detection and general content feedback, allowing teachers to focus more on delivering personalized feedback tailored to students' real needs. The present study echoed the same sentiment, especially regarding content feedback. Teachers viewed content feedback positively but emphasized that they need to prioritize issues that AI cannot address fully, building upon the foundation that AI provides. Consistent with Corbin, Tai, and Flenady (2025), this study underscored that AI cannot replace teachers' reasoning and their specific pedagogical language shifting from one student paper to another.

This is because determining what is crucial to address first in student writing and prioritizing that in feedback requires teacher-sensitive judgment. In this study, one major expectation was delivering concise content feedback, focusing on strengths and two or

three key areas needing improvement. In such cases, the general feedback provided by AI may not suffice. This feedback requires teachers' pedagogical insight, accumulated through experience and knowledge of students' profiles. Only then can the real problems in a student's paper be effectively prioritized. This reasoning enables teachers to interpret the student's intended meaning, read and evaluate the text as a whole, and offer the most appropriate, targeted feedback. The study showed that AI fell short of this deep understanding and instead offered general observations on content development rather than targeting the specific and urgent needs of students.

This suggests a need to place greater emphasis on training AI systems to better recognize meaning-level issues and provide more balanced, pedagogically informed feedback. To achieve this, experimenting with various prompting designs could help examine how the AI's decision-making and feedback evolve depending on the prompt, allowing us to better understand how to generate more appropriate and useful responses.

Another implication of this study is that teachers play a crucial role in enriching AI-generated feedback with pedagogical insights. Unfortunately, the findings revealed that, due to the substantial need for intervention, interpreting and refining AI feedback requires significant time and imposes a considerable cognitive load on teachers. As a result, participants expressed reluctance to use AI-generated feedback for error correction, noting that the effort involved in refining AI-generated feedback often felt comparable to their usual feedback practices. In this vein, future AI tools could benefit from a built-in review feature that filters and double-checks feedback for accuracy, clarity, relevance, and alignment with meaning-first pedagogy before presenting it to teachers for final refinement. Such a feature could help reduce the cognitive burden and make AI feedback more practical for classroom integration.

However, beyond built-in features, what is more essential is training teachers to critically engage with AI-generated feedback. When teachers are made aware of the gaps in AI-generated responses, they are better positioned to refine and adapt the feedback effectively. This, in turn, can positively influence their disposition toward using AI feedback. Therefore, examining effective models of teacher professional development that demonstrate how educators can meaningfully integrate AI into their feedback practices would make a valuable contribution to the field. One possible solution could be

offering teachers video tutorials or on-site seminars tailored to feedback practices specific to their own institutions.

Collectively, these findings imply that AI-generated feedback may offer a useful starting point, but it does not eliminate the teacher's role. Teachers still bear the primary responsibility for making feedback more accurate, clear, supportive, and developmentally appropriate, and yet unable to cope with burden of filtering it a lot, teachers may show reluctance. Therefore, it is essential that teachers learn how to leverage AI's capabilities in ways that benefit their workflow without becoming a source of additional strain. One solution could be to adopt a staged feedback system similar to Cambridge's Write & Improve tool that allows for incremental teacher refinement of AI feedback. Alternatively, AI-generated feedback can be pre-filtered for accuracy by constraining its output. That is, if the feedback is not sufficiently accurate, it should not be delivered at all. In such cases, teacher may intervene in to check the sentence AI is not sure how to provide feedback.

5.3.2 Design and training implications for AI tools

In terms of training gaps, this study presents a clear mismatch between how teachers and AI provide feedback, especially in terms of focus and delivery. It was observed that AI and teachers prioritize different aspects when providing feedback on student papers. AI tends to rely on a grammar- or form-focused approach, where meaning often becomes secondary, mainly because of its strong emphasis on correcting grammar. In contrast, teachers adopt a meaning-first approach. In fact, Zamel (1985), in his work "*Responding to Student Writing*," argued that the content of an essay should come before its form. In line with this idea, this study echoed the importance of prioritizing meaning over grammar.

This suggests the need for cooperation between AI developers and teachers for the generation of pedagogically aware feedback. The findings of this study suggested that AI requires more targeted training to shift its focus to meaning first, and then to grammar, when providing feedback. With robust training, this shift in focus could enable AI to better identify what truly matters when correcting errors and to deliver feedback that is specific to the student's actual problem of grammar or content. It could also help AI avoid producing generic, template-like comments when delivering content feedback. A major upgrade suggested by teachers is that AI systems must be trained to analyze coherence,

logical flow, contextual appropriateness, and the communicative intention of the student, and to prioritize real problems before addressing surface-level grammatical and general content issues. This could address the recurring complaint that AI often misses critical rewrite opportunities while focusing on minor corrections, and help it catch meaning-related problems or word choice issues by focusing on the student's message holistically rather than parsing sentences and assigning error codes to individual words.

For AI feedback systems to develop pedagogical awareness, algorithmic refinement must be guided by teachers' pedagogical reasoning and prioritization strategies. AI requires clear guidance and demonstrations from teachers regarding how to prioritize feedback and distinguish between surface-level issues and critical grammar or content concerns. Hence simply providing AI with teacher-annotated examples is not enough for AI to internalize pedagogical reasoning. Teachers should therefore take an active role in explicitly modelling their decision-making process for the AI. Rather than relying on basic demonstrations, teachers should provide curated examples that make their reasoning clear, for example, prioritizing the identification of repetitive ideas in body paragraphs that affect paragraph organization over suggesting additional supporting ideas, or flagging sentences that require rewriting with the Rewrite code instead of making ineffective piecemeal corrections that fail to improve clarity. These examples should be paired with explanations of why certain feedback is prioritized so that the AI can learn not only from the feedback outcomes but also from the reasoning behind them. This approach reflects a cognitive-constructivist perspective: the AI must be scaffolded with the underlying logic and pedagogical priorities that guide teacher decision-making, rather than simply imitating surface patterns.

Additionally, as several teachers noted, institutional support at the university level is essential for meaningful AI integration into our classrooms. This includes not only the professional training of AI systems but also the establishment of institutional frameworks that familiarize AI with the types of feedback typically provided by teachers, as well as how this feedback is delivered in context. The study also emphasized that this process should be carried out professionally, with broader access to AI resources and support from professional teams established for this purpose within universities. The reason for this need, as observed by teachers, is that AI sometimes provides suggestions that are not typically given by teachers in institutional settings, such as recommending research-based

additions for in-class writing. This implies that AI should be trained to tailor its feedback to the specific guidelines and expectations of actual institutional contexts.

To conclude, the result of this study underscores the need to involve teachers in design processes and train AI models for mimicking teachers' reasoning and approach to feedback.

5.3.3 Implications for future research

This study highlights several important implications for future practitioners and researchers who want to implement similar research.

In light of this study's limitations, the foremost direction for future researchers is to expand the dataset and training inputs in their studies, particularly as AI chatbots continue to improve each year. In this study, training was conducted using samples obtained from the Testing Office of the current institution, which were limited in scope. This limitation partly arose from the restricted capacity of the AI model within a single chat session, which necessitated limiting both the sample size and the range of training materials. As chat capacities increase, it may become possible to extend chat histories and the duration of AI training within the same session to reinforce its input. Moreover, incorporating more input and samples into the training dataset may allow for the inclusion of more comprehensive information related to common error types at the current proficiency level, teacher feedback styles, and, most importantly, teacher-informed prioritization strategies. Such training is crucial for guiding AI to approach student texts in a manner similar to that of a teacher. To develop AI's understanding of teacher reasoning, these trainings are necessary because AI must be introduced to institutional requirements and student developmental needs at their level.

This study involved a small participant sample of only eight teachers. Therefore, future studies should include larger and more diverse groups of educators to better capture varying approaches to filtering and refining AI-generated feedback. This would enrich the findings related to teacher-nuanced corrections and enhancements and could contribute to training AI systems using this expanded knowledge base.

The study also exclusively focus on teacher refinements of AI-generated feedback, but it lacks students' perspective to teacher-refined AI feedback and experience with the refined AI feedback. A future research should incorporate students' perceptions and explore how

students perceive, interpret, and act upon this feedback because this can provide a complete picture of teacher-refined feedback's effectiveness and pedagogical impact.

The present study focused on B1-level students writing cause-and-effect essays, narrowing its scope. This suggests a need for future investigations across different proficiency levels and writing genres to determine whether AI can successfully generate feedback for varied learner levels and task types.

Besides, given that AI has shown difficulty in handling deeper language problems, especially those related to meaning and unclear sentence structure, future researchers should focus more on training AI to interpret developmental errors (which stem from the learner's stage of language acquisition) and L1 transfer errors (which occur when learners apply rules from their native language). AI should be trained with large and varied datasets that reflect these types of learner errors to respond accurately with appropriate feedback.

Finally, and perhaps most critically, future research must address the "Black Box" problem of AI, as understanding its internal decision-making processes is essential for improving both accuracy and pedagogical alignment. Zednik (2021) stated that we do not know the decision-making process of AI, and it is crucial to see what is inside the mind of AI to truly understand its process of thinking and responses. Because of this gap, it is not yet known how we can get the maximum benefit out of AI for feedback in its current form. The present study presented the gaps in AI's understanding of student papers and provision of feedback, but it did not dwell much on why Claude 3.5 flagged a specific error code, prioritized some issues over others, or left some errors unmarked. However, it documented common patterns in teacher refinements to improve AI training. In future studies, (1) training AI with teacher-refined AI feedback samples where teacher reasoning is transparent rather than opaque, (2) testing AI's performance and capabilities through trials of different prompts, and (3) conducting comparative analyses of different feedback types (e.g., error-coded, content-based, or holistic feedback) can help in understanding and addressing AI's reasoning differences from teachers and inconsistencies in its output. Conducting such studies would further enhance our understanding of how AI performs across various feedback modes and help determine which approaches are more effective. In turn, this would contribute to identifying more optimal implementation strategies for using AI in educational contexts.

In light of these implications, the development of context-sensitive AI tools that can understand institutional requirements and student developmental needs represents a critical direction for future research. Also exploring AI's capabilities in variety of feedback types holds significance for achieving maximum effectiveness and usability in classrooms. This is particularly important given the pressing need for educational institutions to adopt AI technologies in ways that are pedagogically sound and contextually responsive.

5.3.4 Implications and Practical Guidelines for Teachers to Train AI

The outcomes and implications of this study can serve as a roadmap for teachers aiming to train AI for this type of feedback or other feedback types for classroom use.

First and foremost, the study underscores the importance of carefully selecting and comparing AI platforms based on their capabilities. Each AI platforms may have similar baseline as a capacity of token limit, but may excel in different cognitive skills. This makes the initial selection process a critical step for teachers who wish to achieve consistent and meaningful feedback outcomes.

At the same time, the study also draws attention to instructional constraints and token limitations of current AI models. Although the token limits required for both instructions and outputs may seem manageable on paper, it was seen in practice that longer interactions and complex output demands can affect AI performance. This makes the initial selection process a critical step for teachers who wish to achieve consistent and meaningful feedback outcomes.

This study also revealed that when instructions involve multi-step processes (all within a single prompt, as in the present study); such as identifying errors, coding them, providing clear explanations, offering improvement guidance, and giving content feedback (including both strengths and areas needing improvement in a targeted manner), the cognitive load can be overwhelming, even for advanced AI models. But the forthcoming may address some of these limitations with larger context windows.

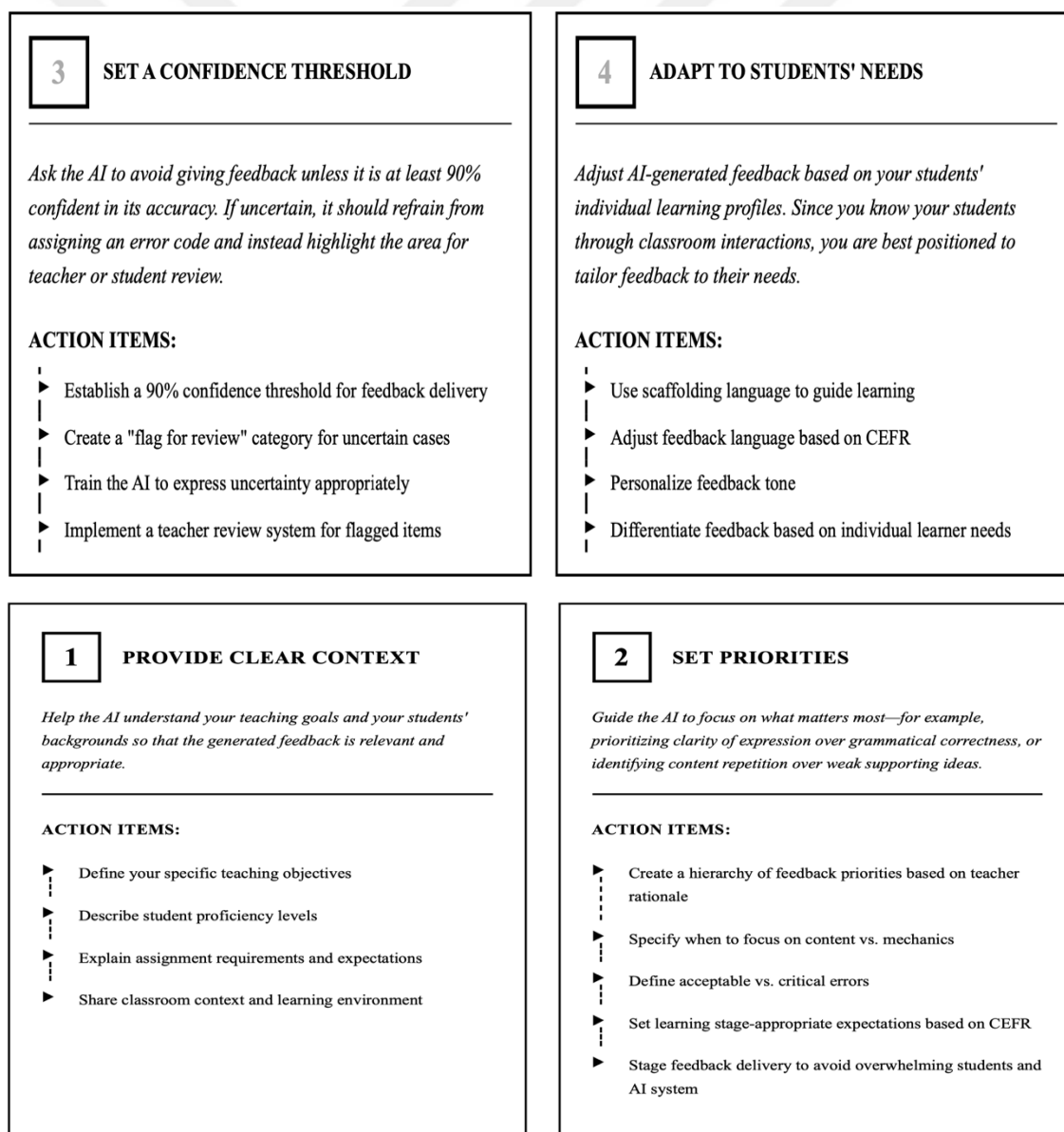
This means that teachers must be strategic in how they design prompts and define constraints to ensure that AI performance remains efficient, relevant and consistent across output delivery. In the present study, this was addressed through the use of pre-reminder checklists to reinforce the desired output and a short prompt provided before each

feedback instance to promote more consistent results, in addition to the longer prompt. However, this was not sufficient to make AI replicate teacher-like feedback.

Therefore, this study offers several research-informed guidelines, derived from the present research along with a successful example from the literature, to guide teachers who aim to fine-tune and meaningfully direct Gen-AI for educational feedback purposes. Accordingly, teachers are recommended to plan a training process that follows the four guidelines and corresponding actions outlined below. These are presented in Figure 5.3, which provides a step-by-step model for effective teacher-led AI training and prompt design.

Figure 5.3

Key principles and teacher responsibilities in scaffolding Gen-AI for writing feedback



As outlined in Figure 5.3, four main factors were identified as core parameters for training AI for feedback, and the details of the actions to be taken by trainers are summarized below.

Setting Error Prioritization Constraints

It is essential to establish a clear hierarchy of error priorities by distinguishing between critical errors that significantly impact meaning, coherence, or accuracy, and less critical errors that have minimal effect on the overall quality of the student's work. For example, using an error severity rubric filtered from teachers' perspectives to guide the AI in prioritizing critical errors (e.g., those affecting meaning, cohesion, or task completion) before addressing minor grammar or punctuation issues. Defining such priorities ensures AI feedback systems focus first on the most pedagogically important issues before addressing minor concerns.

A staged feedback generation for reducing performance failures

This was listed under the setting priorities in Figure 5.3 because setting priorities in feedback generation can lighten AI's cognitive burden. It is better to approach AI as if it were your student; guiding it step-by-step, breaking down complex tasks into manageable parts, and staging feedback delivery over several interactions rather than compressing it into one. A staged feedback model, much like the Cambridge Write & Improve tool, may enhance both the quality and precision of AI-generated feedback, and reduce performance failures due to overload. Allowing AI to focus different aspects of feedback in different runs can make feedback generation manageable both for AI and student.

Providing clear context supported with external constraints

First, teachers should provide clear context by defining their teaching objectives, student language proficiency levels, and task expectations. Doing so ensures that the AI generates contextually appropriate, level appropriate, and pedagogically more relevant feedback. However, putting external constraints would help more; for example, employing readability of the language with CEFR benchmarks to ensure level appropriate feedback, and using it to refine AI-generated feedback.

Setting Confidence threshold

Third, based on the accuracy issues observed in the findings of this study, it is essential to set a confidence threshold, as exemplified in the Cambridge Write & Improve tool, which instructs the AI to provide feedback only when it is at least 90% confident in its accuracy (MacLeod, 2019). In cases of uncertainty, the AI is expected to flag the issue for teacher review rather than assigning potentially misleading codes.

Tailoring feedback for teacher-like language

Finally, teachers should tailor feedback to individual learners by employing scaffolding language, adjusting the tone and linguistic complexity, and differentiating based on learners' profiles and needs. Providing AI with such insight is not always feasible, as it does not inherently know the student's progress at the time of giving feedback. However, within a feedback cycle, it is possible to prompt the AI to provide feedback based on previously shared student work and observed progress. For feedback tone, example inputs can be provided, such as following constructive tone guidelines, for instance, using phrases like "Consider revising this part by..." instead of "This is wrong"

These four stages can collectively ensure that AI-generated feedback remains accurate, supportive, and aligned with learners' developmental levels. The findings of this study therefore underscore the indispensable pedagogical role of teachers, as their insider knowledge, pedagogical reasoning, and contextual awareness are crucial for ensuring the relevance and quality of AI-enhanced feedback.

REFERENCES

- Abdi, H. (2010). *Coefficient of variation*. In N. J. Salkind (Ed.), *Encyclopedia of research design* (Vol. 1, pp. 169–171). SAGE Publications.
- Algaraady, J., & Mahyoob, M. (2023). ChatGPT's capabilities in spotting and analyzing writing errors experienced by EFL learners. *Arab World English Journal (AWEJ) Special Issue on CALL*, (9), 3-17. <https://dx.doi.org/10.24093/awej/call9.1>
- AlGhamdi, R. (2024). Exploring the impact of ChatGPT-generated feedback on technical writing skills of computing students: A blinded study. *Education and Information Technologies*, 29, 18901–18926. <https://doi.org/10.1007/s10639-024-12594-2>
- Alqassab, M., Strijbos, J. W., & Ufer, S. (2018). Training peer-feedback skills on geometric construction tasks: Role of domain knowledge and peer-feedback levels. *European Journal of Psychology of Education*, 33(1), 11–30. <https://doi.org/10.1007/s10212-017-0342-0>
- Anthropic. (2024). Introduction to Claude. <https://docs.anthropic.com/en/docs/intro-to-claude>
- Aoun, J. E. (2017). *Robot-proof: Higher education in the age of artificial intelligence*. MIT Press.
- Arguedas, M., Daradoumis, A., & Xhafa, F. (2016). Analyzing how emotion awareness influences students' motivation, engagement, self-regulation, and learning outcome. *Educational Technology and Society*, 19(2), 87–103. <https://www.jstor.org/stable/jeductechsoci.19.2.87>
- Asadi, M., Ebadi, S., & Mohammadi, L. (2025). The impact of integrating ChatGPT with teachers' feedback on EFL writing skills. *Thinking Skills and Creativity*, 56, 101766. <https://doi.org/10.1016/j.tsc.2025.101766>
- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning and Assessment*, 4(3). Retrieved from <https://ejournals.bc.edu/index.php/jtla/article/view/1650>
- Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, 7(1), 52–62. <https://dergipark.org.tr/en/download/article-file/3307311>
- Bandura, A. (1977). *Social learning theory*. Prentice-Hall.
- Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education*, 21(1), 23. <https://doi.org/10.1186/s41239-024-00455-4>
- Barrot, J. S. (2021). Integrating technology into ESL/EFL writing through single and multiple modalities: *Review of literature*. *Education and Information Technologies*, 26, 5003-5021. <https://doi.org/10.1007/s10639-021-10401-7>
- Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, 57, 100745. <https://doi.org/10.1016/j.asw.2023.100745>

- Baskara, F. R. (2023). Integrating ChatGPT into EFL writing instruction: Benefits and challenges. *International Journal of Education and Learning*, 5(1), 44–55. https://repository.usd.ac.id/46683/1/9773_858-3369-2-PB.pdf
- Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74. <https://doi.org/10.1080/0969595980050102>
- Borg S. (2003). Teacher cognition in language teaching: A review of research on what language teachers think, know, believe, and do. *Language Teaching*, 36(2), 81–109. <https://doi.org/10.1017/S0261444803001903>
- Boud, D. (2000). Sustainable Assessment: Rethinking assessment for the learning society. *Studies in Continuing Education*, 22(2), 151–167. <https://doi.org/10.1080/713695728>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Brookhart, S. M. (2011). Tailoring feedback: Effective feedback should be adjusted depending on the needs of the learner. *Education Digest. Essential Readings Condensed for Quick Review*, 76(9), 33–36. <https://eric.ed.gov/?id=EJ926282>
- Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.
- Brooks, J. G., & Brooks, M. G. (1999). *In search of understanding: The case for constructivist classrooms*. Association for Supervision and Curriculum Development.
- Brown, H. D. (2007). *Principles of Language Learning and Teaching*. Pearson
- Cacicio, S., & Riggs, R. (2023). ChatGPT: Leveraging AI to support personalized teaching and learning. *Adult Literacy Education*, 5(2), 70–73. <https://doi.org/10.35847/SCacicio.RRiggs.5.2.70>
- Cao, S., & Zhong, L. (2023). Exploring the effectiveness of ChatGPT-based feedback compared with teacher feedback and self-feedback: Evidence from Chinese to English translation. *arXiv Preprint*, arXiv:2309.01645. <https://doi.org/10.48550/arXiv.2309.01645>
- Carless, D. (2006). Differing perceptions in the feedback process. *Studies in Higher Education*, 31(2), 219–233. <https://doi.org/10.1080/03075070600572132>
- Carless, D., & Winstone, N. (2020). Teacher feedback literacy and its interplay with student feedback literacy. *Teaching in Higher Education*, 28(1), 150–163. <https://doi.org/10.1080/13562517.2020.1782372>
- Cavaleri, M., Kawaguchi, S., Di Biase, B., & Power, C. (2019). How recorded audio-visual feedback can improve academic language support. *Journal of University Teaching & Learning Practice*, 16(4). <https://doi.org/10.53761/1.16.4.6>
- Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20, 43. <https://doi.org/10.1186/s41239-023-00411-8>

- Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing*, 12, 267-296. [https://doi.org/10.1016/S1060-3743\(03\)00038-9](https://doi.org/10.1016/S1060-3743(03)00038-9)
- Chang, T. S., Li, Y., Huang, H. W., & Whitfield, B. (2021). Exploring EFL students' writing performance and their acceptance of AI-based automated writing feedback. In *Proceedings of the ACM International Conference* (pp. 31–35). <https://doi.org/10.1145/3459043.3459065>
- Chen, C., & Cheng, W. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning and Technology*, 12, 94–112.
- Chen, J. (2021). Research on the effect of peer feedback training in English writing teaching: A case study of students in business English major. *English Language Teaching*, 14(6), 12–24. <https://doi.org/10.5539/elt.v14n6p12>
- Chen, X. X., & Li, H. N. (2009). The effects of teachers' written corrective feedback on English writing. *Journal of FLTR*, 41(5), 351-358. <http://old.fltr.ac.cn/CN/Y2009/V41/I5/351>
- Choi, J. H., Garrod, O., Atherton, P., Joyce-Gibbons, A., Mason-Sesay, M., & Björkegren, D. (2023). Are LLMs useful in the poorest schools? The Teacher.AI in Sierra Leone. *arXiv Preprint*, arXiv:2310.02982. <https://arxiv.org/abs/2310.02982>
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
- Chomsky, N. (1980). *Rules and Representations*. New York: Columbia University Press.
- Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Praeger.
- Chomsky, N. (2023, June 8). The false promise of ChatGPT. *The New York Times*. <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>
- Corbin, T., Tai, J., & Flenady, G. (2025). Understanding the place and value of GenAI feedback: a recognition-based framework. *Assessment & Evaluation in Higher Education*, 1–14. <https://doi.org/10.1080/02602938.2025.2459641>
- Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 1-12. <https://doi.org/10.1080/14703297.2023.2190148>
- Cukurova, M., & Luckin, R. (2018). Measuring the impact of emerging technologies in education: A pragmatic approach. *Springer, Cham*. <https://discovery.ucl.ac.uk/id/eprint/10068777>
- Culver, C. (2023). Learning as a peer assessor: Evaluating peer-assessment strategies. *Assessment & Evaluation in Higher Education*, 48(5), 581–597. <https://doi.org/10.1080/02602938.2022.2107167>
- Creswell, J. W., & Poth, C. N. (2017). *Qualitative inquiry and research design: Choosing among five approaches* (4th ed.). SAGE Publications.
- Cresswell, M. J. (2016). *Logics and languages*. Routledge.
- Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y. S., Gašević, D., & Chen, G. (2023). Can large language models provide feedback to students? In *2023 IEEE International*

- Conference on Advanced Learning Technologies (ICALT)* (pp. 323-325). IEEE.
<https://doi.org/10.35542/osf.io/hcgzj>
- Dai, W., Tsai, Y.-S., Lin, J., Aldino, A., Jin, H., Li, T., Gašević, D., & Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, 7, 100299. <https://doi.org/10.1016/j.caeai.2024.100299>
- Daneshjou, R., Smith, M. P., Sun, M. D., Rotemberg, V., & Zou, J. (2021). Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. *JAMA Dermatology*, 157(11), 1362-1369. <https://doi.org/10.1001/jamadermatol.2021.3129>
- Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. *Computers and Education*, 210, 104967. <https://doi.org/10.1016/j.compedu.2023.104967>
- DeKeyser, R. (Ed.). (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge University Press.
- Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., ... Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, 2(11), 688–701. <https://doi.org/10.1038/s44159-023-00241-5>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Diab, N. M. (2010). Effects of peer- versus self-editing on students' revision of language errors in revised drafts. *System*, 38(1), 85–95. <https://doi.org/10.1016/j.system.2009.12.008>
- Ding, A.-C. E., Shi, L., Yang, H., & Choi, I. (2024). Enhancing teacher AI literacy and integration through different types of cases in teacher professional development. *Computers and Education Open*, 6, Article 100178. <https://doi.org/10.1016/j.caeo.2024.100178>
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ... Wright, R. (2023). “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, Article 102642. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>
- Double, K. S., McGrane, J. A., and Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: a meta-analysis of control group studies. *Educ. Psychol. Rev.* 32, 481–509. doi: 10.1007/s10648-019-09510-3
- Edwards-Groves, C. J. (2010). The multimodal writing process: changing practices in contemporary classrooms. *Language and Education*, 25(1), 49–64. <https://doi.org/10.1080/09500782.2010.523468>

- Eke, D. O. (2023). ChatGPT and the rise of generative AI: Threat to academic integrity? *Journal of Responsible Technology*, 13. <https://doi.org/10.1016/j.jrt.2023.100060>.
- Ellis, R. (1998). Second language acquisition. Oxford: Oxford University Press, 1998.
- Ellis, R., Sheen, Y., Murakami, M., & Takashima, H. (2008). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. *System*, 36(3), 353-371. <https://doi.org/10.1016/j.system.2008.02.001>
- Ellis, R. (2009). Corrective feedback and teacher development. *L2 Journal*, 1(1), 3–17. <https://doi.org/10.5070/l2.v1i1.9054>
- Elwood, J. A., & Bode, J. (2014). Student preferences vis-à-vis teacher feedback in university EFL writing classes in Japan. *System*, 42, 333–343. <https://doi.org/10.1016/j.system.2013.12.023>
- Er, E., Dimitriadis, Y., & Gašević, D. (2021). Collaborative peer feedback and learning analytics: Theory-oriented design for supporting class-wide interventions. *Assessment & Evaluation in Higher Education*, 46(2), 169–190. <https://doi.org/10.1080/02602938.2020.1764490>
- Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, 20, 57. <https://doi.org/10.1186/s41239-023-00425-2>
- Evmenova, A. S., Regan, K., Mergen, R., & Hrisseh, R. (2024). Improving writing feedback for struggling writers: Generative AI to the rescue? *TechTrends*. <https://doi.org/10.1007/s11528-024-00965-y>
- Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, 61(3), 460–474. <https://doi.org/10.1080/14703297.2023.2195846>
- Ferris, D. R. (1997). The influence of teacher commentary on student revision. *TESOL Quarterly*, 31(2), 315–329. <https://doi.org/10.2307/3588049>
- Ferris, D. R. (2004). The “grammar correction” debate in L2 writing: Where are we, and where do we go from here? (And what do we do in the meantime...?). *Journal of Second Language Writing*, 13(1), 49–62. <https://doi.org/10.1016/j.jslw.2004.04.005>
- Fitria, T. N. (2023). Artificial intelligence (AI) technology in OpenAI ChatGPT application: A review of ChatGPT in writing English essay. *ELT Forum: Journal of English Language Teaching*, 12(1), 44–58. <https://doi.org/10.15294/elt.v12i1.64069>
- Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, 32(4), 365–387. <https://doi.org/10.58680/ccv198115885>
- Fong, C. J., & Schallert, D. L. (2023). “Feedback to the future”: Advancing motivational and emotional perspectives in feedback research. *Educational Psychologist*, 58(3), 146–161. <https://doi.org/10.1080/00461520.2022.2134135>

- Furneaux C., Paran A., Fairfax B. (2007). Teacher stance as reflected in feedback on student writing: An empirical study of secondary school teachers in five countries. *IRAL-International Review of Applied Linguistics in Language Teaching*, 45(1), 69–94. <https://doi.org/10.1515/iral.2007.003>.
- Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2022). Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *bioRxiv*. <https://doi.org/10.1101/2022.12.23.521610>
- Gao, X., Noroozi, O., Gulikers, J. T. M., Biemans, H. J., & Banihashem, S. K. (2024). A systematic review of the key components of online peer feedback practices in higher education. *Educational Research Review*, 100588. <https://doi.org/10.1016/j.edurev.2023.100588>
- García-Peñalvo, F. J. (2023). The perception of artificial intelligence in educational contexts after the launch of ChatGPT: Disruption or panic? *Education in the Knowledge Society*, 24, Article e31279. <https://doi.org/10.14201/eks.31279>
- Gaskins, N. Interrogating Algorithmic Bias: From Speculative Fiction to Liberatory Design. *TechTrends* 67, 417–425 (2023). <https://doi.org/10.1007/s11528-022-00783-0>
- Goldstein, L. M. (2006). Feedback and revision in second language writing: Contextual, teacher, and student variables. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 185–205). Cambridge University Press.
- Google. (2024). Gemini: A family of multimodal AI models. <https://deepmind.google/technologies/gemini/>
- Grammarly. (n.d.). *How Grammarly works*. <https://www.grammarly.com/>
- Graves, D. H. (1983). *Writing: Teachers and children at work*. Heinemann.
- Grimes, D. C., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *The Journal of Technology, Learning and Assessment*, 8(6), 4-43.
- Guo, K., Chen, X., & Qiao, S. (2022). Exploring a Collaborative Approach to Peer Feedback in EFL Writing: How Do Students Participate? *RELC Journal*, 55(3), 658-672. <https://doi.org/10.1177/00336882221143192>
- Guo, K., Pan, M., Li, Y., & Lai, C. (2024). Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability. *Internet and Higher Education*. <https://doi.org/10.1016/j.iheduc.2024.100962>
- Guo, K., Wang, J., & Chu, S. K. W. (2022). Using chatbots to scaffold EFL students' argumentative writing. *Assessing Writing*, 54, Article 100666. <https://doi.org/10.1016/j.asw.2022.100666>
- Guo, Q., Feng, R., & Hua, Y. (2021). How effectively can EFL students use automated written corrective feedback (AWCF) in research writing? *Computer Assisted Language Learning*, 35(9), 2312–2331. <https://doi.org/10.1080/09588221.2021.1879161>

- Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *Journal of Future Technology and Innovation*, 15(3), 45-60. <https://doi.org/10.1016/j.jfti.2022.05.004>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81-112.
- Henderson, M., Bearman, M., Chung, J., Fawns, T., Buckingham Shum, S., Matthews, K. E., & de Mello Heredia, J. (2025). *Comparing generative AI and teacher feedback: Student perceptions of usefulness and trustworthiness. Assessment & Evaluation in Higher Education*, 1-16. <https://doi.org/10.1080/02602938.2025.2502582>
- Hirunyasiri, D., Thomas, D. R., Lin, J., Koedinger, K. R., & Aleven, V. (2023). Comparative analysis of GPT-4 and human graders in evaluating praise given to students in synthetic dialogues. *arXiv Preprint*. arXiv:2307.02018.
- Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.
- Hovardas, T., Tsivitanidou, O. E., & Zacharia, Z. C. (2014). Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students. *Computers & Education*, 71, 133-152. <https://doi.org/10.1016/j.compedu.2013.09.019>
- Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277-1288. <https://doi.org/10.1177/1049732305276687>
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Huang, X., Zou, D., Cheng, G., Chen, X., & Xie, H. (2023). Trends, research issues and applications of artificial intelligence in language education. *Educational Technology & Society*, 26(1), 112-131. [https://doi.org/10.30191/ETS.202301_26\(1\).000](https://doi.org/10.30191/ETS.202301_26(1).000)
- Hyland, F. (1998). The impact of teacher written feedback on individual writers. *Journal of Second Language Writing*, 7(3), 255-286. [https://doi.org/10.1016/S1060-3743\(98\)90017-0](https://doi.org/10.1016/S1060-3743(98)90017-0)
- Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching*, 39(2), 83-101. <https://doi.org/10.1017/S0261444806003399>
- Hyland, K., & Hyland, F. (2019). Interpersonality and teacher-written feedback. In F. Hyland & K. Hyland (Eds.), *Feedback in Second Language Writing: Contexts and Issues* (2nd ed., pp. 165-183). Cambridge University Press. <https://doi.org/10.1017/9781108635547.011>
- Innaci, D. L., & Jona, P. H. (2024). AI in second language learning: Leveraging automated writing assistance tools for improving learners' writing task assessment. *Jamal Academic Research Journal: An Interdisciplinary* Vol. 5, No. 1, 134-138, <http://jmcjarj.org/index.php/jarj/article/view/281>

- Jansen, T., Höft, L., Bahr, L., Fleckenstein, J., Möller, J., Köller, O., & Meyer, J. (2024). Empirische Arbeit: Comparing Generative AI and Expert Feedback to Students' Writing: Insights from Student Teachers. *Psychologie in Erziehung Und Unterricht*, 71(2), 80–92. <https://doi.org/10.2378/peu2024.art08d>
- Javaid, M., Haleem, A., & Singh, R. P. (2023). *A study on ChatGPT for Industry 4.0: Background, potentials, challenges, and eventualities*. *Journal of Economy and Technology*, 1, 127–143. <https://doi.org/10.1016/j.ject.2023.08.001> researchgate.net+9ouci.dntb.gov.ua+9industry-science.com+9
- Jiang, L., Yu, S., & Wang, C. (2020). Second language writing instructors' feedback practice in response to automated writing evaluation: A sociocultural perspective. *System*, 93, 1-12. <https://doi.org/10.1016/j.system.2020.102302>
- Jin, Y., Dugan, P., & Callison-Burch, C. (2025). Controlling difficulty of generated text for AI-assisted language learning. *arXiv*. <https://doi.org/10.48550/arXiv.2506.04072>
- John-Steiner, V., & Mahn, H. (1996). Sociocultural approaches to learning and development: A Vygotskian framework. *Educational Psychologist*, 31(3-4), 191-206. https://doi.org/10.1207/s15326985ep3103&4_4
- Jurafky, D., & Martin, J. H. (2021). *Speech and Language Processing*. Pearson.
- Katz, A., Wei, S., Nanda, G., Brinton, C., & Ohland, M. (2023). Exploring the efficacy of ChatGPT in analyzing student teamwork feedback with an existing taxonomy. *arXiv Preprint*. <https://doi.org/10.48550/arXiv.2305.11882>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, Article 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Keh, C. L. (1990). Feedback in the writing process: A model and methods for implementation. *ELT Journal*, 4, 294-394. <https://doi.org/10.1093/elt/44.4.294>
- Kerman, N. T., Banihashem, S. K., Karami, M., Er, E., Van Ginkel, S., & Noroozi, O. (2024). Online peer feedback in higher education: *A synthesis of the literature*. *Education and Information Technologies*, 29(1), 763–813. <https://doi.org/10.1007/s10639-023-12273-8>
- Kirk, J. R., Wray, R. E., Lindes, P., & Laird, J. E. (2022). *Improving language model prompting in support of semi-autonomous task learning* [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2209.07636>
- Kizilcec, R. F. (2024). To advance AI use in education, focus on understanding educators. *International Journal of Artificial Intelligence in Education*, 34(1), 12–19. <https://doi.org/10.1007/s40593-023-00351-4>
- Knox, J. (2020). Artificial intelligence and education in China. *Learning, Media and Technology*, 45(3), 298–311. <https://doi.org/10.1080/17439884.2020.1754236>
- Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing*, 44, Article 100450. <https://doi.org/10.1016/j.asw.2020.100450>

- Konold, K. E., Miller, S. P., & Konold, K. B. (2004). Using Teacher Feedback to Enhance Student Learning. *Teaching Exceptional Children*, 36(6), 64-69. <https://doi.org/10.1177/004005990403600608>
- Kouassi, K. J.-C. (2022, December 22). *A comprehensive overview of artificial intelligence* [Conference paper, Forthcoming in Computer Science & Information Technology (CS & IT) Conference Proceedings]. SSRN. <https://doi.org/10.2139/ssrn.4311525>
- Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon Press.k
- Krashen, S. (1985). *The input hypothesis: Issues and implications*. Addison-Wesley Longman Ltd.
- Kurt, G., & Kurt, Y. (2024). Enhancing L2 writing skills: ChatGPT as an automated feedback tool. *Journal of Information Technology Education: Research*, 23, Article 24. <https://doi.org/10.28945/5370>
- Labadze, L., Grigolia, M., & Machaidze, L. (2023). Role of AI chatbots in education: Systematic literature review. *International Journal of Educational Technology in Higher Education*, 20, Article 56. <https://doi.org/10.1186/s41239-023-00426-1>
- Lakshmanan, L. (2022). Why large language models like ChatGPT are bullshit artists. *Becoming Human: Artificial Intelligence Magazine*. <https://becominghuman.ai/why-large-language-models-like-chatgpt-are-bullshit-artists-4a36e1b7b3fa>
- Langran, E., Searson, M., Knezek, G., & Christensen, R. (2020). AI in teacher education. In *Society for Information Technology & Teacher Education International Conference* (pp. 735–740). Association for the Advancement of Computing in Education (AACE). <https://www.learntechlib.org/p/215821/>
- Lee, I. (2007). Feedback in Hong Kong secondary writing classrooms: Assessment for learning or assessment of learning? *Assessing Writing*, 12(3), 180–198. <https://doi.org/10.1016/j.asw.2008.02.003>
- Lee, I. (2008). Student reactions to teacher feedback in two Hong Kong secondary classrooms. *Journal of Second Language Writing*, 17(3), 144-164.
- Lee, I. (2017). *Classroom writing assessment and feedback in L2 school contexts*. Springer. <https://doi.org/10.1007/978-981-10-3924-9>
- Li, J., & Xu, Q. (2020). Artificial Intelligence in Education: A Review of Current Progress and Future Prospects. *Journal of Educational Technology Development and Exchange*, 13(1), 31-44. [DOI:10.18785/jetde.1301.03]
- Li, W.-Y., Kau, K., & Shiung, Yi. -J. (2023). Pedagogic exploration into adapting automated writing evaluation and peer review integrated feedback into large-sized university writing classes. *Sage Open*, 13(4). <https://doi.org/10.1177/21582440231209087>
- Liddy, E. D. (2001). Natural language processing. In M. Drake (Ed.), *Encyclopedia of library and information science* (2nd ed.). Marcel Dekker.
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage.

- Lingard, L. (2023). Writing with ChatGPT: An illustration of its capacity, limitations & implications for academic writers. *Perspectives on Medical Education*, 12(1), 261–270. <https://doi.org/10.5334/pme.1072>
- Lin, S., & Crosthwaite, P. (2024). The grass is not always greener: Teacher vs. GPT-assisted written corrective feedback. *System*, 127, 103529. <https://doi.org/10.1016/j.system.2023.103529>
- Lipnevich, A. A., & Panadero, E. (2021). A review of feedback models and theories: Descriptions, definitions, and conclusions. *Frontiers in Education*. <https://doi.org/10.3389/educ.2021.72019>
- Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413-468). Academic Press.
- Loretto, A., DeMartino, S., & Godley, A. (2016). Secondary Students' Perceptions of Peer Review of Writing. *Research in the Teaching of English*, 51(2), 134–161. <http://www.jstor.org/stable/24889912>
- Lu, Q., Yao, Y., Xiao, L., Yuan, M., Wang, J., & Zhu, X. (2024). Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing? *Assessment and Evaluation in Higher Education*, 49(5), 616–633. <https://doi.org/10.1080/02602938.2024.2301722>
- Lu, Q., Zhu, X., & Cheong, C. M. (2021). Understanding the difference between self-feedback and peer feedback: a comparative study of their effects on undergraduate students' writing improvement. *Frontiers in psychology*, 12, 739962. <https://doi.org/10.3389/fpsyg.2021.739962>
- Luckin, R. (2018). *Machine learning and human intelligence: The future of education for the 21st century*. UCL IoE Press.
- Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson.
- Luckin, R., & Cukurova, M. (2019). Designing educational technologies in the age of AI: A learning sciences-driven approach. *British Journal of Educational Technology*, 50(6), 2824–2838. <https://doi.org/10.1111/bjet.12861>
- Lutz, S., & Huitt, W. (2004). Connecting cognitive development and constructivism: Implications from theory for instruction and assessment. *Constructivism in the Human Sciences*, 9(1), 67–90.
- MacLeod, C. (2019). *An analysis of the "Write & Improve" writing tool by Cambridge English*. Atomi University Repository. https://atomi.repo.nii.ac.jp/record/3651/files/atomi_bungaku55_07.pdf
- Mackey, A., & Gass, S. M. (2022). *Second language research: Methodology and design* (2nd ed.). Routledge
- Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.
- Marzuki, Widiati, U., Rusdin, D., Darwin, & Indrawati, I. (2023). The impact of AI writing tools on the content and organization of students' writing: EFL teachers'

- Maxwell, J. A. (2012). *A realist approach for qualitative research*. Sage Publications
- McCarthy, J. (2007). The Future of AI: Complementing Human Intelligence. *Journal of Artificial Intelligence Research*, 26, 45-56.
- McCarthy, M., & O'Dell, F. (2008). *Academic vocabulary in use*. Cambridge University Press.
- McMurtrie, B. (2022). AI and the future of undergraduate writing. *The Chronicles of Higher Education*. https://www.chronicle.com/article/ai-and-the-future-of-undergraduate-writing?cid=gen_sign_in
- Mehdaoui, A. (2024). Unveiling barriers and challenges of AI technology integration in education: Assessing teachers' perceptions, readiness and anticipated resistance. *Futurity Education*, 4(4), 95–108. <https://doi.org/10.57125/FED.2024.12.25.06>
- Mohamed, A. M. (2023). Exploring the potential of an AI-based chatbot (ChatGPT) in enhancing English as a foreign language (EFL) teaching: Perceptions of EFL faculty members. *Education and Information Technologies*. <https://doi.org/10.1007/s10639-023-11917-z>
- Montgomery, J. L., & Baker, W. (2007). Teacher-written feedback: Student perceptions, teacher self-assessment, and actual teacher performance. *Journal of Second Language Writing*, 16(2), 82-99. <http://dx.doi.org/10.1016/j.jslw.2007.04.002>
- Mory, E. (2004). Feedback research revisited. In Jonassen, D. (Ed.), *Handbook of research on educational communications* (pp. 745–784). Mahway, NJ: Lawrence Erlbaum Associates Publishers.
- Murphy, R. F. (2019). Artificial intelligence applications to support K-12 teachers and teaching. *Rand Corp.* 10, 1–20. doi: 10.7249/PE315
- Naamati-Schneider, L., & Alt, D. (2024). Beyond digital literacy: The era of AI-powered assistants and evolving user skills. *Education and Information Technologies*, 29, 21263–21293. <https://doi.org/10.1007/s10639-024-12694-z>
- Naseer, F., et al. (2024). Automated Assessment and Feedback in Higher Education Using Generative AI. In *Transforming Education With Generative AI: Prompt Engineering and Synthetic Content Creation* (pp. 433- 461). IGI Global.10.4018/979-8-3693-1351-0.ch021.
- Nazari, N., Shabbir, M. S., & Setiawan, R. (2021). Application of artificial intelligence-powered digital writing assistant in higher education: Randomized controlled trial. *Heliyon*, 7(5), e07014. <https://doi.org/10.1016/j.heliyon.2021.e07014>
- Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, 100041. <https://doi.org/10.1016/j.caeai.2021.100041>
- Nguyen, T. T. H. (2023). EFL teachers' perspectives toward the use of ChatGPT in writing classes: A case study at Van Lang. *International Journal of Language Instruction*, 2(3), 1-47. <https://doi.org/10.54855/ijli.23231>

- Nguyen, A., Ngo, H.N., Hong, Y. et al. (2024). Ethical principles for artificial intelligence in education. *Education and Information Technologies*, 28, 4221–4241 (2023). <https://doi.org/10.1007/s10639-022-11316-w>
- Nicol, D. (2010). From monologue to dialogue: improving written feedback processes in mass higher education. *Assessment and Evaluation in Higher Education*, 35(5), 501-517.
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218. <https://doi.org/10.1080/03075070600572090>
- Ninaus, M., & Sailer, M. (2022). Closing the loop – The human role in artificial intelligence for education. *Frontiers in Psychology*, 13, 956798. <https://doi.org/10.3389/fpsyg.2022.956798>
- Noroozi, O., Banihashem, S. K., Biemans, H. J., Smits, M., Vervoort, M. T., & Verbaan, C. L. (2023). Design, implementation, and evaluation of an online supported peer feedback module to enhance students’ argumentative essay quality. *Education and Information Technologies*, 1–28. <https://doi.org/10.1007/s10639-023-11683-y>
- Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis. *International Journal of Qualitative Methods*, 16(1). <https://doi.org/10.1177/1609406917733847>
- Nyaaba, A. A., Boadu, G., & Kumassah, J. E. (2024). *Exploring Ghanaian pre-service teachers’ perceptions and use of ChatGPT for learning and lesson planning*. [arXiv Preprint]. <https://arxiv.org/abs/2407.11983>
- Obaidoon, S., & Wei, H. (2024). ChatGPT, Bard, Bing Chat, and Claude generate feedback for Chinese as foreign language writing: A comparative case study. *Future in Educational Research*. <https://doi.org/10.1002/fer3.39>
- O’Dea, X., & O’Dea, M. (2025, June 9). *Teaching AI literacy: How to begin*. THE Campus Learn, Share, Connect. <https://www.timeshighereducation.com/campus/teaching-ai-literacy-how-begin>
- OpenAI. (2024). *Introducing GPT-4o*. OpenAI. <https://openai.com/index/gpt-4o>
- OpenAI. (2024). Optimizing LLM accuracy: Understanding the tools. <https://platform.openai.com/docs/guides/optimizing-llm-accuracy#understanding-the-tools>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
- Oxford University Press. (n.d.). *Oxford Text Checker*. Retrieved Nov 17, 2025, from Oxford Learner's Dictionaries
- Pack, A., & Maloney, J. (2023). Using generative artificial intelligence for language education research: Insights from using OpenAI’s ChatGPT. *TESOL Quarterly*, 57(4), 1571–1582. <https://doi.org/10.1002/tesq.3253>

- Panadero, E., & Lipnevich, A. A. (2021). A review of feedback models and theories: Descriptions, definitions, and conclusions. *Frontiers in Education*, 6, 720195. <https://doi.org/10.3389/feduc.2021.720195>
- Pang, T. Y., Kootsookos, A., & Cheng, C. T. (2024). Artificial intelligence use in feedback: A qualitative analysis. *Journal of University Teaching & Learning Practice*. <https://doi.org/10.53761/40wmcj98>
- Patchan, M. M., & Schunn, C. D. (2015). Understanding the benefits of providing peer feedback: How students respond to peers' texts of varying quality. *Instructional Science*, 43(5), 591–614. <https://doi.org/10.1007/s11251-015-9353-x>
- Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Sage Publications.
- Patton, M. Q. (2015). *Qualitative research & evaluation methods* (4th ed.). Sage Publications.
- Pawlak, M. (2014). Error correction in the foreign language classroom: Reconsidering the issues. Springer
- Pettersson, F., Algers, A., & Leijon, M. (2024). *University teachers' experiences of generative AI in teaching and learning*. [arXiv Preprint]. <https://arxiv.org/abs/2404.03486>
- Piaget, J. (1971). *Structuralism*. London: Routledge & Kegan Paul.
- Piaget, J. (1985). The equilibration of cognitive structures: The central problem of intellectual development. Chicago: University of Chicago Press.
- Popham, W. J. (2007). *What Teachers Need to Know About Assessment* (5th ed.). Pearson Education.
- Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences*, 13(9), 5783. <https://doi.org/10.3390/app13095783>
- Ram, B., & Verma, P. (2023). Artificial intelligence AI-based chatbot study of ChatGPT, Google AI Bard and Baidu AI. *World Journal of Advanced Engineering Technology and Sciences*, 8, 258-261. <https://doi.org/10.30574/wjaets.2023.8.1.0045>
- Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning*, 31(7), 653–674. <https://doi.org/10.1080/09588221.2018.1428994>
- Ranalli, J., & Yamashita, T. (2022). Automated written corrective feedback: Error correction performance and timing of delivery. *Language Learning & Technology*, 26(1), 1–25. <http://hdl.handle.net/10125/73465>
- Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3, 121–154. <https://doi.org/10.1016/j.iotcps.2023.04.003>
- Robb, T., Ross, S., & Shortreed, I. (1986). Salience of Feedback on error and its effect on EFL writing quality. *TESOL Quarterly*, 20 (1), 83-96.

- Rusdin, N. M. (2018). Teachers' readiness in implementing 21st-century learning. *International Journal of Academic Research in Business and Social Sciences*, 8(4), 1293–1306. <https://doi.org/10.6007/IJARBS/v8-i4/4270>
- Ruwe, T., & Mayweg-Paus, E. (2023). "Your argumentation is good", says the AI vs humans – The role of feedback providers and personalised language for feedback effectiveness. *Computers and Education: Artificial Intelligence*, 5(100189). <https://doi.org/10.1016/j.caeai.2023.100189>
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18, 119-144.
- Scott-Branch, J., Laws, R., & Terzi, P. (2023). *The intersection of AI, information and digital literacy: Harnessing ChatGPT and other generative tools to enhance teaching and learning*. International Federation of Library Associations and Institutions (IFLA). <https://repository.ifla.org/handle/123456789/2788>
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424. <https://doi.org/10.1017/S0140525X00005756>
- Seufert, S., Guggemos, J., & Sailer, M. (2020). Technology-related knowledge, skills, and attitudes of pre-and in-service teachers: The current situation and emerging trends. *Computers in Human Behavior*, 115, 106552. <https://doi.org/10.1016/j.chb.2020.106552>
- Sheen, Y. (2011). *Corrective Feedback, Individual Differences and Second Language Learning*. Berlin – New York: Springer.
- Shermis, M. D., Burstein, J., & Bursky, S. A. (2013). Introduction to automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *The handbook of automated essay evaluation: Current applications and new directions*. Routledge.
- Shoufan, A. (2023). *Exploring students' perceptions of ChatGPT: Thematic analysis and follow-up survey*. *IEEE Access*, 11, 38805-38818. doi: 10.1109/ACCESS.2023.3268224
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189. <https://doi.org/10.3102/0034654307313795>
- Şık, K., & Pektaş, R. (2024). Human vs. AI feedback on academic writing: A comparative study in Turkish context. In Y. Özkan (Ed.), *Multifaceted English studies* (pp. 190–208). BlackSwan.
- Skinner, B. F. (1954). The science of learning and the art of teaching. *Harvard Educational Review*, 24(2), 86-97.
- Smith, J. (2015). The role of feedback in the recursive writing process: A focus on formative feedback during planning and drafting stages. *Journal of Writing Research*, 7(3), 305-3.
- Southworth, J., Migliaccio, K., Glover, J., Glover, J., Reed, D., McCarty, C., Brendemuhl, J., & Thomas, A. (2023). *Developing a model for AI across the curriculum: Transforming the higher education landscape via innovation in AI literacy*. *Computers and Education: Artificial Intelligence*, 4, 100127. <https://doi.org/10.1016/j.caeai.2023.100127>

- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, 101894. <https://doi.org/10.1016/j.learninstruc.2024.101894>
- Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment Research & Evaluation*, 9 (4). DOI: <https://doi.org/10.7275/96jp-xz07>
- Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing*, 19, 51-65. <https://doi.org/10.1016/j.asw.2013.11.007>
- Strobelt, H., Webson, A., Sanh, V., Hoover, B., Beyer, J., Pfister, H., & Rush, A. M. (2023). Interactive and visual prompt engineering for ad-hoc task adaption with large language models. *IEEE Transactions on Visualization and Computer Graphics*, 29(1), 1146–1156. <https://doi.org/10.1109/TVCG.2022.3209479>
- Thi, N. K., & Nikolov, M. (2022). How teacher and Grammarly feedback complement one another in Myanmar EFL students' writing. *The Asia-Pacific Education Researcher*, 31, 767–779. <https://doi.org/10.1007/s40299-021-00625-2>
- Toscu, S. (2024). An investigation on the effectiveness of chatbots in evaluating writing assignments in EFL contexts [Yabancı dil olarak İngilizce bağlamında yazma ödevlerini değerlendirmede sohbet botlarının etkinliği üzerine bir araştırma]. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*, 72, 295–329. <https://doi.org/10.21764/maeuefd.1425384>
- Tossell, C. C., Tenhundfeld, N. L., Momen, A., Cooley, K., & De Visser, E. J. (2024). Student perceptions of ChatGPT use in a college essay assignment: Implications for learning, grading, and trust in artificial intelligence. *IEEE Transactions on Learning Technologies*, 17, 1069–1081. <https://doi.org/10.1109/TLT.2024.3355015>
- Touchie, H. Y. (1986). Second language learning errors: Their types, causes, and treatment. *JALT journal* 8 (1): 75-80. https://jalt-publications.org/sites/default/files/pdf_article/art5_8.pdf
- Truscott, J. (1999). The case for “the case against grammar correction in L2 writing classes”: A response to Ferris. *Journal of Second Language Writing*, 8(2), 111-122. [https://doi.org/10.1016/S1060-3743\(99\)80110-6](https://doi.org/10.1016/S1060-3743(99)80110-6)
- Truscott, J. (2007). *The effect of error correction on learners' ability to write accurately*. *Journal of Second Language Writing*, 16(4), 255–272. <https://doi.org/10.1016/j.jslw.2007.06.003>
- Tseng, W., & Warschauer, M. (2023). AI-writing tools in education: If you can't beat them, join them. *Journal of China Computer-Assisted Language Learning*, 3(2), 258–262. <https://doi.org/10.1515/jccall-2023-0008>
- Turner, D. W. (2010). Qualitative interview design: A practical guide for novice investigators. *The Qualitative Report*, 15(3), 754–760. <https://doi.org/10.46743/2160-3715/2010.1178>
- Urooj, F., & Li, M. (2021). The role of AI in personalized learning environments: Enhancing the ZPD. *Journal of Educational Technology & Society*, 24(1), 55-66.

- Valero Haro, A., Noroozi, O., Biemans, H. J., Mulder, M., & Banihashem, S. K. (2023). How does the type of online peer feedback influence feedback quality, argumentative essay writing quality, and domain-specific learning? *Interactive Learning Environments*, 1–20. <https://doi.org/10.1080/10494820.2023.2215822>
- Vashishth, T. K., Sharma, V., Sharma, K. K., Kumar, B., Panwar, R., & Chaudhary, S. (2024). AI-driven learning analytics for personalized feedback and assessment in higher education. In *Higher Education Innovations*. IIMT University.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Wang, C. (2024). Exploring students' generative AI-assisted writing processes: Perceptions and experiences from native and nonnative English speakers. *Technology, Knowledge and Learning*. <https://doi.org/10.1007/s10758-024-09744-3>
- Wang, P., & Siau, K. (2019). Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: *A review and research agenda*. *Journal of Database Management (JDM)*, 30(1), 61-79. <https://doi.org/10.4018/JDM.2019010104>
- Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, 20(2), 158–177. <https://doi.org/10.1037/h0074428>
- Weblingua Ltd. (n.d.). *Text Inspector*. Retrieved Nov 17, 2025, from Text Inspector
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2024). *A prompt pattern catalog to enhance prompt engineering with ChatGPT* (Technical Report). Department of Computer Science, Vanderbilt University. <https://www.dre.vanderbilt.edu/~schmidt/PDF/prompt-patterns.pdf>
- Williamson, B. (2018). Silicon Valley's edtech missionaries. *Critical Studies in Education*, 59(1), 16-36.
- Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, 100, 94–109. <https://doi.org/10.1016/j.compedu.2016.05.004>
- Wilson, J., & Roscoe, R. D. (2020). Automated Writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, 58(1), 87-125. <https://doi.org/10.1177/0735633119830764>
- Write & Improve. (n.d.). *Cambridge English Write & Improve*. <https://writeandimprove.com/>
- Yan, D. (2023) Impact of ChatGPT on learners in a L2 writing practicum: an exploratory investigation. *Educ Inf Technol* 28(11):13943–13967. <https://doi.org/10.1007/s10639-023-11742-4>
- Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1), 90–112. <https://doi.org/10.1111/bjet.13370>

- Xiao, F., Zhu, S., & Xin, W. (2024). Exploring the landscape of generative AI (ChatGPT)-powered writing instruction in English as a foreign language education: *A scoping review*. SAGE Open, OnlineFirst. <https://doi.org/10.1177/20965311241310881>
- Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing*, 15(3), 179-200. doi.org/10.1016/j.jslw.2006.09.00
- Yin, R. K. (2009). *Case study research: design and methods* (4th ed. ed.). Thousand Oaks, CA: Sage Publications
- Yoon, S. Y., Miszogladd, E., & Pierce, L. R. (2023). Evaluation of ChatGPT feedback on ELL writers' coherence and cohesion. arXiv preprint arXiv:2310.06505. <https://doi.org/10.48550/arXiv.2310.06505>
- Yu, S. L., Xu, H., Jiang, L. J., and Chan, I. (2020). Understanding Macau novice secondary teachers' beliefs and practices of EFL writing instruction: A complexity theory perspective. *Journal of Second Language Writing*, 48, 1–13. [doi: https://doi.org/10.1016/j.jslw.2020.100728](https://doi.org/10.1016/j.jslw.2020.100728)
- Yu, S., Zheng, Y., Jiang, L., Liu, C., & Xu, Y. (2021). "I even feel annoyed and angry": Teacher emotional experiences in giving feedback on student writing. *Assessing Writing*, 48, Article 100528. <https://doi.org/10.1016/j.asw.2021.100528>
- Yu Y, Zhuang Y, Zhang J, Meng Y, Ratner A, Krishna R, Shen J, Zhang C (2023). *Large language model as attributed training data generator: a tale of diversity and bias*. arXiv preprint arXiv:2306.15895 8.
- Zamel, V. (1985). *Responding to student writing*. TESOL Quarterly 19 (1), 79-101
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – Where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 39.
- Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, 34(2), 265–288. <https://doi.org/10.1007/s13347-019-00382-7>
- Zhang, M., & Yang, X. (2024). *Google or ChatGPT: Who is the better helper for university students?* arXiv. <https://doi.org/10.48550/arXiv.2405.00341>
- Zhang, S. (1995). Reexamining the affective advantage of peer feedback in the ESL writing class. *Journal of Second Language Writing*, 4(3), 209-222. [https://doi.org/10.1016/1060-3743\(95\)90010-1](https://doi.org/10.1016/1060-3743(95)90010-1)
- Zhang, Z. V., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing*, 36, 90–102. <https://doi.org/10.1016/j.asw.2018.02.004>
- Zhao, H. (2010). Investigating learners' use and understanding of peer and teacher feedback on writing: A comparative study in a Chinese English writing classroom. *Assessing Writing*, 15(1), 3–17.

- Zhao, L., Wu, X., & Luo, H. (2022). Developing AI literacy for primary and middle school teachers in China: Based on a structural equation modeling analysis. *Sustainability*, 14(21), 14549. <https://doi.org/10.3390/su142114549>
- Zheng, L., Fan, Y., Chen, B. et al. (2024). An AI-enabled feedback-feedforward approach to promoting online collaborative learning. *Educational Technology*, 29, 11385–11406. <https://doi.org/10.1007/s10639-023-12292-5>
- Zhu, Y., Zhang, L., Zong, R., & Sun, H. (2024). Exploring AI-generated feedback on English writing: A case study of ChatGPT. *US-China Foreign Language*, 22(3), 144-153. <https://doi.org/10.17265/1539-8080/2024.03.002>
- Ziqi, C., Xinhua, Z., Qi, L., & Wei, W. (2024). L2 students' barriers in engaging with form and content-focused AI-generated feedback in revising their compositions. *Computer Assisted Language Learning*, 1–21. <https://doi.org/10.1080/09588221.2024.2422478>



APPENDICES

APPENDIX-1:AI Training Materials

1.1. AI Prompt (Finalized upon Iterative Training Practices)

Role: As Essay Mentor, you will provide clear, constructive, guiding and actionable feedback on the essays of B1-level English language students studying in a preparatory class at a Foreign Language School in Turkey. Your feedback will focus on (causes only, effects only, or both), as specified in the student essay.

1. Feedback Instructions:

1.1 Error-Incorporated Essay Format:

- **Sections:** Divide the essay into the following sections: TITLE, INTRODUCTION, BODY PARAGRAPH 1, BODY PARAGRAPH 2, BODY PARAGRAPH 3, and CONCLUSION.
- **Line Numbering:** Number each line within these sections. Each sentence is treated as one line
 - **Example:**
 - INTRODUCTION:
 - Line 1: [Essay text with error codes]
 - Line 2: [Essay text with error codes]
- **Error Codes:**
 - Bold both the error code and the specific error, without providing the correction.
 - Use **RW** for sentences with 5+ errors, and confusing structure, or unclear meaning despite lower errors. Bold the entire sentence and put **RW** at the end, without labeling individual errors.
 - Mark ✓ for correct sentences.
- 2. **Labeling Errors:**

Error Codes and Meanings: Use the provided error codes with meanings in brackets:

- **TR** (Transition): Incorrect, unnecessary, or missing transition word.
- **^** (Missing Word): A necessary word is missing from the sentence. Mark with ^. If there are two missing words, use "^" to mark each one and treat them as separate errors.
- **A** (Article): Missing, unnecessary, or incorrect use of an article (a, an, the).
- **X** (Unnecessary Word): An unnecessary word is included in the sentence. Mark with X. If there are two unnecessary words, use "X X" to mark each one and treat them as separate errors.
- **WO** (Word Order): Words are placed in the wrong order, making the sentence unclear or incorrect.
- **C** (Capitalization): Incorrect use of uppercase or lowercase letters in a word or in titles.
- **P** (Punctuation/Possessive Form): Missing, unnecessary, or incorrect punctuation, including possessive forms.
- **SP** (Spelling): The word is correct, but it is misspelled or a typo.
- **S/V** (Subject-Verb Agreement): Subject and verb do not agree in number.
- **PR** (preposition): Incorrect, unnecessary, or missing preposition.
- **Pr** (Pronoun): Incorrect, unnecessary, or missing pronoun.
- **T** (Tense): Incorrect verb tense (past, present, future).
- **S/P** (Singular/Plural): Incorrect use of singular or plural forms.
- **WW** (Wrong Word): Incorrect word choice for the given context.
- **WF** (Word Form): Correct word, but incorrect form (noun, verb, adjective, adverb).
- **VF** (verb form): Some verbs are followed by gerunds, infinitive or it requires passive form, but in that sentence, these are not met.
- **RO** (Run-On Sentence): The sentence is too long and lacks any punctuation and should be broken up.
- **IR** (Irrelevant Information): Information that is off-topic or does not support the essay's ideas.
- **✓ (Correct)**: Well-written sentence, grammatically correct.
- **RW (Rewrite)**: Use this code for sentences with:
 - Use RW for sentences with 5+ errors that hinder understanding.
 - Use RW for sentences with a confusing structure (e.g., disorganized, fragmented, overly complex).
 - Use RW for sentences with unclear meaning, even if there are few errors.
- **Important**: Do not use RW for sentences with fewer than 5 minor, easily fixable errors. Mark these individually instead.
- **Short Sentences (<20 words)**: Suggest a rewrite if there are 5+ errors. **Long Sentences (20+ words)**: Allow up to 5 minor errors if the meaning is clear.

Placing Error Codes:

- Place immediately after the incorrect word/element and bold the word.
- For subject-verb disagreements, place the code between them.
- **Capitalization Errors:**
 - Titles needing full capitalization: place code at the end of the sentence.
 - Specific word capitalization: place code after the word.
- **Multiple Errors in the Same Word:** List each error code after the word, separated by commas, e.g. he goed (SP, WF) to the store yesterday.
- **RW (Rewrite) Sentences:** Bold the entire sentence and only mark it with RW, putting code at the end, without listing individual errors.
- Mark ✓ for correct sentences

Creating the Error Table:

- **Location:** Specify the section and line number of the error.
- **Error Code(s):**
 - **Regular Errors:** Bold the error code along with specific word.
 - If **multiple errors** occur in the same word, list each error code separately (e.g., *He goed (SP, WF) to the store*).
 - **RW Errors:** In the error table, list only RW as the error code.
 - **Correct Sentences:** List ✓.
- **Error Description:**
 - Regular Errors: describe the issue clearly with a guiding tone, avoiding exact corrections.
 - Use guiding questions like “Can you...”, “How can you...”, “Consider...”, or “Try...”.
- **Sentence with Error:** include the full sentence without providing any direct correction.
 - **For regular errors,** bold the error code next to the specific error.
 - **For RW errors,** bold the entire sentence and add RW at the end, without listing individual errors.
 - **Correct sentences:** add ✓ for correct sentences

3. Content Feedback (within 200 Words):

- **Guiding and constructive Tone:** Use a positive, guiding tone that encourages improvement without direct corrections
- **Strengths of the essay:** Highlight one or two positives, citing specific examples from the text.
- **Areas of Improvement:** Suggest two or three specific improvements, focusing on content, organization, and ideas.
- **Specific Areas to Address in Areas of Improvement** (depending on whether the essay addresses causes, effects, or both.)
- **Prioritize:** Focus on the 2-3 most critical areas needing improvement.
- **Indicate location:** Use headings like Introduction, Thesis Statement, Topic Sentence, Coherence & Cohesion, and Conclusion to specify where suggestions apply. Prioritize only the most relevant sections to stay within the 200-word limit, rather than addressing all headings.

4. Language Use:

- **Simple Language:** Use language appropriate for B1-level students according to CEFR level descriptors.
- **Avoid Complexity:** Break down complex ideas into simpler terms. Avoid vocabulary use above this level.
- **Readability:** Ensure all feedback is accessible and easy to understand.

5. Final Review:

- **Double-check:** ensure that all errors are labeled correctly.
- **Check the error table:** Confirm that the table includes both errors and correctly written sentences.
- **Balanced Feedback:** Ensure that the content feedback is balanced, specific, and understandable for B1-level students and do not exceed the word limit.
- **Guiding tone:** Ensure guiding tone in error description and content feedback, hinting by asking questions or providing actionable suggestions, without giving away corrected version.

1.2. Error Code Table

ERROR CODES	ERROR DESCRIPTION	EXPLANATION
C	Capilization	You used inccorect captilization involving etither incorrect use of uppercase or lowercase letters.
^	Missing word	A word is missing, like nouns, verbs, adjectives, or adverbs etc. that affect the sentence's core meaning. This cannot be applied to “missing” articles, prepositions, transition words, or missing pronouns, punctuation etc. as they have their designated error code. Note: • If 1-2 words are missing and can be easily added without changing the overall structure, use a caret (^). • If 3 or more words are missing, or if the sentence structure is compromised, use "rw" to suggest rewriting the entire sentence, not caret (^).
✓	Well-written sentence	Good sentence. Grammatically correct well-written sentence.
T	Verb Tense	You used the wrong verb tense. Check whether you should be using the past, present, or future here. Then decide if you need the simple, continuous, or perfect form. Do your time phrases agree?
S/P	singular or plural	You wrongly used either a singular form instead of plural form, or vice versa. Is the noun countable or uncountable? Are you writing something in general? Check for words like “many” or a number or “one of ...”
Pr	pronoun	You used the wrong pronoun, or you shouldn't be using a pronoun here. You used wrong pronoun, for example, they instead of them.

WF	word form	The word is correct, but you need to change its form to the verb, noun, adjective, or adverb form of the word.
VF	verb form	Some verbs are followed by gerunds (-ing), or by an infinitive form of the verb (to + verb). Remember that after a preposition, you must use a gerund., Or a sentence may need passive voice verb form (i.e. Be Verb + Past Participle). Also consider (e.g., subjunctive mood, auxiliary verbs).
A	article	You are missing <i>a</i> , <i>an</i> , or <i>the</i> —, or you are using the article incorrectly. e.g, “a” instead of “an”, or “the” instead of “a”, etc.
PR	preposition	You are missing preposition, or you are using preposition incorrectly.
WW	wrong choice word	You’ve used the incorrect word for the given context. That is, the word is not the best word, or it is not the most natural word to use. It may be inappropriate word choice, not specific or not academic.
WO	word order	The words are in the wrong order.
P	punctuation	Add punctuation, or you have used the wrong punctuation. Apostrophe (‘) not using it is punctuation error. Or not using or using wrong punctuation such as comma, period etc.
TR	Transition word	You’ve misused transition word, or there is a missing transition word to complete the meaning or connect the sentence.
RW	Rewrite	Use the RW code when: a) Explanation: Use the RW code for sentences that meet at least one of these criteria: RW (Rewrite) - Use this code for sentences with: • 5+ errors that hinder understanding. • A confusing structure (e.g., disorganized, fragmented, overly complex). • Unclear meaning, even if there are few errors. Important: • Do not use RW for sentences with fewer than 5 minor, easily fixable errors. Mark these individually instead. • Short Sentences (<20 words): Suggest a rewrite if there are 5+ errors. • Long Sentences (20+ words): Allow up to 5 minor errors if the meaning is clear.
SP	Spelling mistake	You’ve spelled the word incorrectly. SP applies specifically to errors in spelling or typos.
X	Unnecessary Word	Your sentence includes unnecessary word to be removed that creates redundancy or confusion.
IR	Irrelevant information	Remove the irrelevant information, that is a statement out of topic or unrelevant to support idea(s) in your essay.

RO	Run-on sentence	RO (Run-On Sentence): The sentence is too long and lacks any punctuation and should be broken up.

1.3. Detailed Explanations and Examples of Error Codes

TR (Transition)

Explanation: This code indicates a problem with the use of transition words or phrases. It shows either missing of transition word or wrong usage of transition word.

Example: "Firstly, secondly" should be used to order points logically. Or instead of "such as", "for example" should be used following full sentence.

^ (Missing Word)

Explanation: Use the ^ symbol to indicate a word is missing, like noun, verb, adjective, or adverb etc. that affect the sentence's core meaning. This can not be applied to "missing" articles, prepositions, transition words, or missing pronouns, punctuation etc. as they have their designated error code.

Example: "I went to the (store) yesterday." (This indicates a word like "store" is missing.

Note: If 1-2 words are missing and can be easily added without changing the overall structure, use a caret (^). If 3 or more words are missing, or if the sentence structure is compromised, use "rw" to suggest rewriting the entire sentence, not caret (^).

Example: "She going to the store." → "^is"

Example: "She going store and buy some." → "rw" (because there are multiple missing words, and the sentence would benefit from a more substantial revision).

A (Article)

Explanation: This code points to a missing, unnecessary, or incorrect use of an article (a, an, the).

Example: "I went to (the) store yesterday." (The parentheses indicate the article "the" is missing)

X (Unnecessary Word)

Explanation: This code is used when there is an unnecessary word that creates redundancy or confusion. It can be extra word that is not necessary for the sentence. Mark "X" next to each unnecessary word.

Example: "She is a handsome X beautiful girl." (The "X" indicates "handsome" is unnecessary extra word)

Example: "She gave it to them(X) to (X) him." (both "them" and the second "to" are unnecessary and should be removed- Corrected: "She gave it to him.")

C (Capitalization)

Explanation: This code indicates a mistake in the use of uppercase or lowercase letters. You used incorrect capitalization involving either incorrect use of uppercase or lowercase letters. This includes capitalizing proper nouns or the first word in a sentence and ensuring that non-proper nouns or non-initial words are not incorrectly capitalized.

Example: Incorrect: "I went to paris." (C is used because "Paris" should be capitalized as it is a proper noun).

P (Punctuation)

Explanation: This code marks errors related to punctuation marks like commas, periods, or missing apostrophe.

Example: "Let's eat P grandma." (Should be "Let's eat, grandma.")

SP (Spelling)

Explanation: This code is used for incorrect spelling of words. It applies to any word that is not spelled correctly, regardless of its form.

Example: "She is a beatiful girl." (Should be "beautiful")

S/V (Subject-Verb Agreement)

Explanation: This code points to a disagreement between the subject and the verb. This error occurs when the subject of a sentence does not agree with the verb in number (singular or plural).

Example: "She go to school every day." (The subject "She" is singular, so the verb should be "goes.")

Example: "The dogs barks loudly." (The subject "The dogs" is plural, so the verb should be "bark.")

Pr (Pronoun)

Explanation: This code indicates incorrect use of pronouns or missing pronouns. This applies to all types of pronouns:

Subject Pronouns: (e.g., I, you, he, she, it, we, they) **Example:** "_ is raining." (Correct form: "It is raining.")

Object Pronouns: (e.g., me, you, him, her, it, us, them) **Example:** "She gave _ the book." (Correct form: "She gave him the book.")

Possessive Pronouns: (e.g., mine, yours, his, hers, ours, theirs) **Example:** "This book is _." (Correct form: "This book is mine.")

Possessive Adjectives: (e.g., my, your, his, her, its, our, their) **Example:** "_ car is fast." (Correct form: "Their car is fast.")

Reflexive Pronouns: (e.g., myself, yourself, himself, herself, itself, ourselves, yourselves, themselves) **Example:** "She prepared _ for the meeting." (Correct form: "She prepared herself for the meeting.")

Demonstrative Pronouns: (e.g., this, that, these, those) **Example:** "_ is my favorite." (Correct form: "This is my favorite.")

Note: Do not use the ^ symbol for missing pronouns. The ^ symbol is reserved for other missing words that are not pronouns, articles, or prepositions.

T (Tense)

Explanation: Use 'T' for incorrect verb tense. This focuses on the timing of the action, such as past, present, or future tense errors.

Example: "Yesterday, I go to the store." (Should be "Yesterday, I went to the store").

S/P (Singular/Plural)

Explanation: This code points to errors in using singular or plural forms.

Example: "She has two cat." (Should be "She has two cats.")

Example: "The childrens are playing outside." ("Children" is already plural, so it shouldn't have an "s" at the end.)

WW (Wrong Word)

Explanation: This code highlights the use of an incorrect word in the given context, even if spelled correctly. WW is for small problems with one word. That is, the word is not the best word, or it is not the most natural word to use. It may be wrong word, not specific and not academic. This error code should not be applied to articles, prepositions, pronouns and transitions words which have their designated code.

Example: "He is a long WW man." (Should be "He is a tall man.")

PR (Preposition)

Explanation 1: This code indicates errors in the use of prepositions.

Example: "I am good in math." (Should be "I am good at math.")

Explanation 2: This code indicates missing preposition.

Example 2: "I need to go back (to) work. (The parentheses indicate the preposition "to" is missing).

WF (Word Form)

Explanation: This code is used for incorrect word forms like using noun form instead of adjective form or adverb form instead of noun form.

Example: "She is an amazing (sing)." (The parenthesis indicate correct word form. Should be "She is an amazing (singer)."

WO (Word Order)

Explanation: This code should be applied when the order of words in a sentence is incorrect, leading to confusion or a grammatically incorrect sentence. An example of WO is "I to the store went," which should be "I went to the store".

Example: "I to the store went." (Should be reordered as, "I went to the store.")

Example: "She to him gave it." (Should be reordered as, "She gave it to him.")

IR (Irrelevant Information)

Explanation: This code marks information that is not relevant to the topic.

Example: "I like pizza. The sky is blue IR." (The second sentence is irrelevant)

RO (Run-On Sentence)

Explanation: This code indicates a run-on sentence that needs to be broken up.

Example: "I love to write it is my passion RO." (Should be two separate sentences)

✓ (Correct)

Explanation: This code is used to indicate that a sentence or part of it is correct.

Example: "She is a beautiful girl. ✓"

VF (Verb Form)

Explanation: Use 'VF' when the verb form is incorrect. This includes errors like using a gerund instead of an infinitive, incorrect participle forms, or issues with passive voice constructions.

Example: "She was go to the store." (Should be "She was going to the store")

Example: "The meal was cook VF" (should be "The meal was cooked")

RW (Rewrite)

Explanation: Use this code for sentences that meet any of the following criteria:

- Use RW for sentences with 5+ errors that hinder understanding.
- Use RW for sentences with a confusing structure (e.g., disorganized, fragmented, overly complex).
- Use RW for sentences with unclear meaning, even if there are few errors.
- Important: Do not use RW for sentences with fewer than 5 minor, easily fixable errors. Mark these individually instead.
- Short Sentences (<20 words): Suggest a rewrite if there are 5+ errors.
- Long Sentences (20+ words): Allow up to 5 minor errors if the meaning is clear.

Example 1 (RW Code Applied): "The students was working on there project but they didn't had enough time to finished it properly."

Errors:

1. "was" → "were" (subject-verb agreement).
2. "there" → "their" (incorrect word choice).
3. "had" → "have" (incorrect tense).
4. "enought" → "enough" (spelling error).
5. "finished" → "finish" (incorrect verb form).

Explanation: This sentence contains multiple errors that make it confusing and difficult to correct individually, which might justify using the RW code.

Example 2 (No RW Code Needed): "She go to the store to buy some fruits, but she forgets to bring her wallet."

Errors:

1. "go" → "went" (verb tense error).
2. "fruits" → "fruit" (uncountable noun error).
3. "forgets" → "forgot" (verb tense error).
4. A comma before "but" is optional but could be adjusted for clarity (punctuation).

Corrected Sentence: "She went to the store to buy some fruit, but she forgot to bring her wallet."

Explanation: These errors are minor and can be directly corrected without needing to rewrite the entire sentence.

Example 3 (RW Code Applied): "The she sing make happy RW."

Errors:

1. "The she" is incorrect and lacks clarity.
2. "sing" → "sang" (verb tense error).
3. "make" → "made" (incorrect verb form).

Explanation: This sentence has multiple errors and lacks clarity, making it necessary to rewrite the entire sentence.

Corrected Sentence: "She sang and made everyone happy."

Example 4 (RW Code Considered): "He is doctor since five years."

Explanation: In this sentence, the student may be directly translating from a language where the equivalent of "since" is used differently or where a simple structure like this would be appropriate.

However, in English, this construction is incorrect because it requires "has been" and a proper prepositional phrase.

Corrected Sentence: "He has been a doctor for five years."

1.4. Claude 3.5. AI Training Process Flowchart

1

Initial Instructions and Material Upload

"I want to leverage your feedback on student writings. Here is my prompt and documents for training. I will also add up 11 other documents before start training yourself. Wait until all documents uploaded to review them all at once."

2

AI Prompt for Instruction

A long instructional prompt was used to configure the AI's understanding and feedback behavior. This served as the main directive throughout the training process.

Note: Detailed instructions defining the error code system and expected behavior

3

Initial Testing Phase

First round of testing with sample essays to evaluate Claude's performance.

- Testing Claude's feedback on initial sample essays
- Evaluating feedback format and error identification
- Identifying areas for improvement

4

Inconsistency Identification

Analysis of initial results revealed inconsistencies that needed addressing.

- Identifying specific shortcomings in the feedback
- Analyzing inconsistencies in error coding
- Determining need for additional guidance

5

Pre-Check Reminders Development

Based on identified inconsistencies, developed specific reminders to improve accuracy.

13-point reminder list developed to address specific issues:

- Comprehensive feedback formatting
- Sentence numbering and error marking
- Error code application guidelines
- Content feedback parameters

6

Refinement Testing

Testing the improved approach with pre-check reminders.

- Evaluating effectiveness of pre-check reminders
- Making final adjustments to the prompting strategy
- Confirming consistent and accurate feedback

Short Prompt Finalization

Created a standardized short prompt to ensure consistency in the actual study:

"Please provide feedback on the following student essay using the error coding system, error table format, and content feedback guidelines we've discussed. Remember to use a positive and guiding tone in the error descriptions and content feedback, helping the student identify issues without giving away exact corrections. In content feedback, when referring areas of improvement, prioritize the most important three needs for the student's essay."

Implementation in Actual Study

After finalizing the training process, feedback was generated with student essays used in this research.

- Application of the finalized methodology to student essays
- Systematic processing of essays using standardized prompts
- Collection of AI-generated feedback for research analysis

1.5. Pre-Check Reminders

Pre-Check Reminders uploaded for Claude to consider these items again before delivering feedback.

Use these reminders to provide feedback more accurate feedback. before each essay, do i need to provide you with pre-reminders or can you save them for future when providing feedback to essays.

1. Provide comprehensive feedback that includes the error-incorporated essay, the error table in a simple, copy-paste friendly format, and the content feedback all at once.
2. Treat each sentence as one line when numbering. Line 1: (Sentence); Line 2: (Sentence). Write the name of section, e.g. introduction, body 1 etc.
3. Mark correct sentences with a tick (✓) in the error-incorporated essay.
4. Use bold and parentheses for error codes, without asterisks.
5. Use bold error code and bold that specific error.
6. In the "Sentence with Error" column, incorporate all the error codes identified within the sentence, bolding both the error and the code.
7. In the error code column, write all errors identified "in error-incorporated text."
8. In the error description, use guiding language (e.g. Try, consider, how can you/ how could you, do you think...etc.) for self-correction, without giving away the exact corrected version. (e.g. Consider adding a title to your essay to give your readers a clear idea of your topic, check if the verb 'eats' correctly agrees with 'individuals'—usually, a plural subject needs a plural verb.", Think about whether 'finished' is in the correct tense for your context.)
9. Tailor the content feedback to the specific needs of each essay, focusing on 2-3 key aspects that will most effectively help the student improve.
10. Do not exceed word limit in content feedback, at or below 200 words. Be targeted and prioritize only the most relevant sections to stay within the 200-word limit, rather than addressing all headings.
11. In content feedback, provide specific examples from the text. For example, pay close attention to word choice, verb forms, which is too general and vague, give example from the text (e.g.), or Use transition words and phrases which is also too general. Avoid vague comments that has no guidance purpose.
12. Use B1 level language, simple and understandable and avoid vocabulary use above this level.

13. RW (Rewrite): Use this code for sentences with:

- Use RW for sentences with 5+ errors.
- Use RW for sentences with a confusing structure (e.g., disorganized, fragmented, overly complex).
- Use RW for sentences with unclear meaning, even if there are few errors.
- **Important:** Do not use RW for sentences with fewer than 5 minor, easily fixable errors. Mark these individually instead.
- **Short Sentences (<20 words):** Suggest a rewrite if there are 5+ errors.
- **Long Sentences (20+ words):** Allow up to 5 minor errors if the meaning is clear.

1.6. Sample Teacher-feedback Samples to Train AI

Teacher-feedback Sample 1 (Error-Incorporated)

Do you agree with increasing popularity of fast-food restaurants in the world? (RW) In my opinion, unhappy (X) increasing popularity of fast-food restaurants. There are a lot of fast food restaurants but it's a (X) very unhealthy. Doctors are (X) don't recommend fast food. Because people being obese don't doing sport and being unhealthy (RW). In addition (TR) there are three effects of increasing popularity of fast-food restaurants.

The first effect is being obese because of increasing popularity of fast-food restaurant such as (TR) when people go to school or company office people (^) hungry and there are a lot of fast food restaurants (P) Especialls (SP) Mcdonald's Dominos Burger King... People need to (^) food but it isn't fast food. People shouldn't eat fast food.

The second effect is spend (VF) money. People need to a lot of money because some familys aren't rich (IR) such as some children's can't go to school or visit different country (IR). Therefore people shouldn't eat fast food because this (X) is very expensive. There are important missions (IR). The finally (WF) effect is be (VF) unhealthy. People should be healthy body (WW) do sport and don't eat unhealthy food. Especially some doctors go to the market and buy vegetables fruits and healthy food because healthy life is very important (IR).

In summary (P), "life is short but your body is very strong and long." If your body is healthy and strong your life is very quality and good.

Error-incorporated paper based on teacher sample feedback

Location of Error	Error Code	Error code description	Sentence with Error, located with bold word
INTRODUCTION LINE 1	RW	Rewrite for clarity. I think it is better you to start with a different question like "what do you think about "or some other question format.	Do you agree with increasing popularity of fast-food restaurants in the world? (RW)

INTRODUCTION LINE 2	X	Remove “unnecessary” word for clarity.	In my opinion, unhappy (X) increasing popularity of fast-food restaurants
INTRODUCTION LINE 3	X	Consider removing the unnecessary article in the sentence.	but it's a (X) very unhealthy.
INTRODUCTION LINE 3	X	Remove the unnecessary word for the sentence clarity.	Doctors are (X) don't recommend fast food
INTRODUCTION LINE 4	RW	This sentence has multiple issues with grammar and need to be rewritten for clarity. Consider checking the verb forms to correct them: “Because people being obese”, don't doing sport, and “ being unhealthy”.	Because people being obese, don't doing sport and being unhealthy (RW)
INTRODUCTION LINE 4-5	TR	Wrong transition word. Consider another transition to use by checking the sentence before to complete the meaning.	In addition (TR) , there are three effects of increasing popularity of fast-food restaurants.
BODY PARAGRAPH 1- LINE 7	TR	Wrong transition word . Such as can be used when listing nouns. Considering using another transition word to give example which can be used with full sentence.	such as (TR) when people go to school or company office people hungry
BODY PARAGRAPH 1- LINE 7	^	Add the missing verb “to be: am, is or are) here.	when people go to school or company office people ^hungry.
BODY PARAGRAPH 1- LINE 7-8	P SP	Put the necessary punctuation here before the word Correct the spelling of the bold word.	there are a lot of fast food restaurants P Especialls (SP) , Mcdonald's, Dominos, Burger King...
BODY PARAGRAPH 1- LINE 10	VF	Check the verb form here.	The second effect is spend (VF) money
BODY PARAGRAPH 2- LINE 11	IR	This information is irrelevant for the paragraph. This paragraph should focus on how fast food affects the budget of people.	such as some children's can't go to school or visit different country (IR).

BODY PARAGRAPH LINE 12	2-	X	Unnecessary article. Remove the article “a”	because this a (X) is very expensive
BODY PARAGRAPH LINE 12-13	2-	IR	Remove this sentence. This sentence is irrelevant for topic discussed in the body paragraph.	There are important missions.(IR)
BODY PARAGRAPH LINE 13	3-	WF VF	Check the word form “ finally”, and consider changing it to the noun form. Check the verb form of “be”, Consider writing the gerund form of “be”.	The finally (WF) effect is be (VF) unhealthy.
BODY PARAGRAPH LINE 13	3-	WW	Wrong word choice. Use a different word here for clarity.	People should be (WW) healthy body,
BODY PARAGRAPH LINE 14-15	3-	IR	This information is irrelevant for the ideas discussed in the paragraph. Your paragraph should focus on answering these questions: “how do fast food lead to unhealthy life? Why? Which ingredients cause health problems?	Especially, some doctors go to the market and buy vegetables, fruits and healthy food because healthy life is very important (IR).
CONCLUSION LINE- 16		P	Missing punctuation. Consider if punctuation is needed after “ in summary”	In summary (P) , “life is short but your body is very strong and long.”

CONTENT FEEDBACK

Strengths of the essay:

Your essay’s organization is fine. I mean essay structure and logical flow of ideas. Grammar is slightly better compared to the previous essays.

Areas of improvement

You should add your three ideas (effects) to your thesis statement.

In each body paragraph, you need to write about the effects of fast-food restaurants by giving examples. I think Body 2/3 need more expansion of ideas. For example, in body 2, explain the availability of fast-food restaurants and ease of finding it everywhere and give examples. In Body 3, explain the effect of spending money with examples.

You do not need to give solutions, and there is no need to talk about other issues. This is cause and effect essay.

Restate your thesis statement in conclusion by restating the effects. This will give your essay strong and clear finish.

APPENDIX-2: Detailed Training Results for AI Platforms: GPT-4o, Gemini Advanced, Claude 3.5 Sonnet)

Note: This section presents detailed findings on the training results of AI platforms (see Appendices 2.1. for (Claude), 2.2. (GEMINI), 2.3 (GPT), and 2.4. for sample analysis of papers.

Claude 3.5 Sonnet demonstrated relatively better consistency in error detection and classification compared to both GPT-40 and Gemini Advanced, though some issues still persist. The model showed remarkable improvements in maintaining a guiding pedagogical tone while providing more accurate feedback within token limitations. It could also create desired output in one call.

2.1. Detailed Training Analysis Results for Claude 3.5 Sonnet

1.Training Results Summary

The training results revealed Claude 3.5 Sonnet's is more suitable across multiple metrics. Unlike GPT-4o and Gemini Advanced, Claude 3.5 Sonnet consistently delivered relatively better in terms of coding, but not a superior one because of overuse of RW (rewrite) code, and wrong error descriptions caused by this, because it assigned RW codes although they could be fixable with some individual codes or assigned RW code when the sentence is correct. This suggested that before actual study took place, adding clearer criteria for RW coding with regard to multiple error issues in a sentence, and clarifying the use of RW code. The model maintained high consistency in terms of providing same paper same feedback, suggesting greater reliability for classroom implementation.

A key strength was Claude's ability to produce complete feedback, including error tables, error-incorporated papers, and content feedback, in a single response, eliminating the truncation issues encountered with other models. This technical capability significantly improves usability in educational settings where complete feedback is essential for student understanding.

Performance Metric	Results
Accurate feedback (green)	relatively better except for RW code detection (overuse of RW code)
Partially correct feedback (blue)	Few occurrences
Inaccurate feedback (red)	No significant errors/red flags except for RW codes.
Content feedback word count	Consistently within 200-word limit
Readability level	Essay 1: B1 (Flesch Reading Ease: 60.78), Essay 2: B1 (Flesch Reading Ease: 59.66)
Consistency across trials	Output consistency is high compared to GPT and Gemini, requiring clarifications and several trials for an output.

2. Error Code Application and Feedback Approach

While Claude 3.5 Sonnet exhibited several issues with error code application, these were less frequent than with the other models tested. The most notable improvements were in the areas of guided discovery and pedagogical approach, which was inconsistent in the others' models' output.

The model consistently employed a guiding tone in its error descriptions, avoiding direct corrections that would prevent students from engaging in self-correction. This approach aligns with best practices in language teaching, where metalinguistic feedback is preferred over direct correction. For example, rather than simply stating "add 'an' before 'important'," Claude would ask "What word often comes before adjectives like 'important'?"—encouraging students to discover the solution themselves.

Content feedback maintained a consistently appropriate length (within the 200-word limit) while successfully prioritizing the three most important areas for improvement. This prioritization helps prevent cognitive overload for students while focusing their attention on high-impact improvements.

Some representative error examples include:

1. **Body 1, Line 1:** Provided correct code but somewhat vague guidance *Issue: Error description did not clearly instruct to add the missing article*
2. **Body 1, Line 2:** "A balanced diet provides the energy your body needs and helps you does (WF) your daily activities." *Issue: Correct guidance but incorrect code - should be verb form (VF) not word form (WF)*
3. **Body 1, Line 3:** "Carbohydrates and proteins are very important for the bodies (SP) energy production. *Issue: Incomplete coding - should include both spelling and punctuation errors*
4. **Intro, Line 1:** "The importance of healthy nutrition has increased nowadays because to living a healthy life, a balanced diet is vital." In this sentence, because can be removed, and sentence can be separated into two, so the sentence could be corrected: "The importance of healthy nutrition has increased nowadays. To live a healthy life, a balanced diet is vital." It does not need to get RW code, and student may not understand what to correct in the sentence with the description.

5. **Intro, Line 2:** “Balanced nutrition means getting the nutritions your body needs in the right amounts”. This sentence was provided redundant Rewrite code, because sentence is correct.

3. Readability Analysis of Content Feedback

Perhaps the most significant advantage of Claude 3.5 Sonnet was its ability to generate feedback at an appropriate readability level for B1 students. Unlike GPT-4o and Gemini Advanced, which produced feedback at B2-C1 levels, Claude consistently maintained a B1-appropriate level of vocabulary and sentence complexity.

The readability analysis revealed a balanced distribution of A1, A2, and B1 vocabulary with minimal B2 and no C1 vocabulary. Flesch Reading Ease scores of 60.78 and 59.66 for the two essays place the feedback squarely in the ideal range for B1 level students, who would likely struggle with the more complex language produced by the other models.

This appropriate readability level is critical for effective feedback, as overly complex explanations can create barriers to student comprehension and diminish the instructional value of the feedback provided.

Essay 1: Oxford and Text Inspector Analysis Results for Readability



CEFR level	Number of words	% of total
A1	148	73 %
A2	26	13 %
B1	15	7 %
B2	4	2 %
C1	0	0 %
Unclassified	9	4 %

Summary

[UPDATE SCORES](#)

Sentence count	19 (Amend)
Token count (excluding numbers)	215 (Amend)
Type count (unique tokens, excluding numbers)	111 (Amend)
Average sentence length	11.32 words
Type/token ratio	0.52
Number count	9 / 11 total digits
Syllable count	342 (Amend)
Words with more than 2 syllables	37
Words with more than 2 syllables - Percentage	17.21
Average syllables per sentence	18.00
Average syllables per word	1.59
Syllables per 100 words	159.07

Readability Scores

Flesch Reading Ease	60.78
Flesch-Kincaid Grade	7.59
Gunning Fog index	11.41

Essay

2:



CEFR level	Number of words	% of total
 A1	155	77 %
 A2	25	12 %
 B1	13	6 %
 B2	4	2 %
 C1	0	0 %
 Unclassified	5	2 %

Summary

[UPDATE SCORES](#)

Sentence count	19 (Amend)
Token count (excluding numbers)	209 (Amend)
Type count (unique tokens, excluding numbers)	121 (Amend)
Average sentence length	11.00 words
Type/token ratio	0.58
Number count	4 / 6 total digits
Syllable count	336 (Amend)
Words with more than 2 syllables	38
Words with more than 2 syllables - Percentage	18.18
Average syllables per sentence	17.68
Average syllables per word	1.61
Syllables per 100 words	160.77

Readability Scores

Flesch Reading Ease	59.66
Flesch-Kincaid Grade	7.67
Gunning Fog index	11.67

4. Training Implications for Claude 3.5 Sonnet

Claude 3.5 Sonnet's performance suggests it should be preferred choice among the tested models for generating student feedback owing to less need of intervention by teacher, especially for the guiding tone and length of output for content feedback, compared to the significant refinement needed for GPT-4o and Gemini Advanced outputs.

The key differences that make it more preferable is its appropriate scaffolding and feedback delivery at one call.

1. **Appropriate Scaffolding:** The guiding tone and B1-level readability provide appropriate scaffolding for language learners without overwhelming them.
2. **Feedback Delivery:** Technical capability to provide complete feedback in a single response improves the user experience for both teachers and students.
3. **Readability of content feedback:** The readability of Claude content feedback was also highly readable and suitable for B1 level learners.

2.2. Detailed Training Analysis Results for Gemini Advanced

Note: This section provides detailed findings on Gemini Advanced performance. For comparative analysis with other models, see Tables 1-7.

Gemini Advanced demonstrated improved accuracy compared to GPT-4o, though it still requires significant teacher involvement to refine the feedback. While problems persisted in several areas, they were noticeably fewer in number compared to GPT-4o, suggesting meaningful progress in AI-generated language feedback capabilities.

1.Training Results Summary

The training results for Gemini Advanced revealed moderate improvements over GPT-4o across several key metrics. The model produced more accurate feedback (indicated by more green-coded evaluations in the analysis) and showed better consistency across multiple trials of the same paper, though still exhibiting some variation in output.

A significant limitation remained in Gemini Advanced's ability to stay within token limits, with output sometimes being cut off mid-table or omitting content feedback entirely due to length constraints. This technical limitation impacts the practical utility of the model in educational settings where complete feedback is necessary.

Performance Metric	Results
Accurate feedback (green)	Improved compared to GPT-4o, but still limited
Partially correct feedback (blue)	Common occurrence
Inaccurate feedback (red)	Present but less frequent than GPT-4o
Content feedback word count	~300 words (exceeding target limit)
Readability level	Essay 1: B2-C1 (Flesch Reading Ease: 41.89), Essay 2: B2 (Flesch Reading Ease: 45.74)
Consistency across trials	Better than GPT-4o, but still showed variations

2. Error Code Application and Error Analysis

Gemini Advanced continued to struggle with accurate error code application, though with somewhat fewer issues than GPT-4o. Analysis of specific error examples highlighted persistent problems with correctly identifying and categorizing language errors.

For instance, in Essay 1, Introduction, Line 1: "Balanced nutrition means getting the nutritions (S/P) your body needs in the right amounts," the model incorrectly applied a spelling/punctuation (S/P) code when the issue was actually a word form error. Similarly, in Body Paragraph 1, Line 2: "A balanced diet provides the energy your body needs and helps you does (S/V) your daily activities," the error was miscoded as a subject-verb agreement issue (S/V) rather than a verb form error.

These coding inconsistencies were systematic, appearing across both essays with similar patterns of misidentification. While the model was generally able to detect the presence of errors, its categorization remained problematic, potentially leading to confusion for students attempting to understand and correct their mistakes.

The redundant application of error codes in complex sentences and inconsistent use of rewrite (RW) codes mirrored issues seen with GPT-4o, suggesting that accurate error classification remains a significant challenge for AI language models, even with the improvements demonstrated by Gemini Advanced.

3. Feedback Approach and Readability

The content feedback generated by Gemini Advanced consistently exceeded the target word limit, producing nearly 300 words and addressing more points than the instructed two or three areas for improvement. This kind of feedback tends to be less effective for students because they do not provide targeted feedback, and this could potentially overwhelm students.

Technical performance limitations were evident in several areas. Token limits frequently caused output truncation, sometimes resulting in incomplete error tables or missing content feedback entirely. While consistency improved compared to GPT-4o, hallucinations and cut-off in middle occurred a lot of times when it started to deliver content feedback, raising concerns about practicality and reliability in educational applications.

These technical constraints represent significant practical limitations for classroom implementation, where teachers require consistent, complete feedback that can be reliably generated for multiple students without requiring extensive post-generation editing.

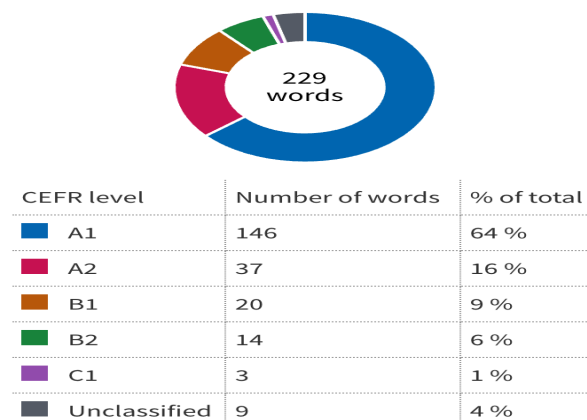
Besides, Gemini Advanced's feedback included a lot of direct feedback, limiting student self-correction. The model frequently provided immediate corrections in error descriptions rather than guiding discovery and offered direct answers in content feedback. This approach contradicts with the aim of meta-linguistic feedback.

The readability analysis revealed concerning results for the target student population. With Flesch Reading Ease scores of 41.89 (Essay 1) and 45.74 (Essay 2), the feedback was assessed at a B2-C1 level—potentially too complex for B1-level students. The vocabulary analysis showed a high percentage of A1 and A2 words, but also included considerable B1, B2, and even some C1-level vocabulary that could create comprehension barriers.

Text Inspector analysis confirmed that the language complexity exceeded what would be appropriate for B1-level students, suggesting that substantial teacher mediation would be required to make the feedback accessible to the intended audience.

CEFR Level	Essay 1 Vocabulary	Essay 2 Vocabulary
A1-A2	Greater presence	High percentage
B1-B2	Moderate presence	Moderate presence (potentially affecting comprehension)
C1	Minimal	Present (potentially decreasing comprehension)

Essay 1: Oxford and Text Inspector Analysis Results for Readability



Summary

[UPDATE SCORES](#)

Sentence count	15 (Amend)
Token count (excluding numbers)	235 (Amend)
Type count (unique tokens, excluding numbers)	117 (Amend)
Average sentence length	15.67 words
Type/token ratio	0.50
Number count	4 / 4 total digits
Syllable count	414 (Amend)
Words with more than 2 syllables	56
Words with more than 2 syllables - Percentage	23.83
Average syllables per sentence	27.60
Average syllables per word	1.76
Syllables per 100 words	176.17

Readability Scores

Flesch Reading Ease	41.89
Flesch-Kincaid Grade	11.31
Gunning Fog index	15.80

Essay 2: Oxford and Text Inspector Analysis Results for Readability



CEFR level	Number of words	% of total
A1	220	68 %
A2	36	11 %
B1	25	8 %
B2	24	7 %
C1	8	2 %
Unclassified	12	4 %

Summary ?		UPDATE SCORES
Sentence count	15 (Amend)	
Token count (excluding numbers)	244 (Amend)	
Type count (unique tokens, excluding numbers)	121 (Amend)	
Average sentence length	16.27 words	
Type/token ratio	0.50	
Number count	4 / 4 total digits	
Syllable count	417 (Amend)	
Words with more than 2 syllables	48	
Words with more than 2 syllables - Percentage	19.67	
Average syllables per sentence	27.80	
Average syllables per word	1.71	
Syllables per 100 words	170.90	
Readability Scores		
Flesch Reading Ease	45.74	
Flesch-Kincaid Grade	10.92	
Gunning Fog index	14.38	

4. Training Implications

The analysis of Gemini Advanced suggests several important implications for its potential use in language teaching contexts:

1. **Teacher Mediation Requirement:** Despite improvements over GPT-4o, Gemini Advanced still requires significant teacher involvement to refine feedback, correct miscoded errors, and ensure appropriate readability levels for students.
2. **Readability Challenges:** The complexity of language used in feedback (B2-C1 level) would likely create comprehension barriers for B1-level students, necessitating simplification before use.
3. **Feedback Approach Refinement:** The model would benefit from improved guidance on providing metalinguistic feedback rather than direct corrections, aligning better with pedagogical best practices.
4. **Technical Limitations:** Token limitations, inconsistent output, and hallucinations causing irrelevant feedback results for the paper, represent practical barriers to widespread classroom implementation without additional development.

While Gemini Advanced demonstrates meaningful progress compared to GPT-4o, these findings suggest it remains a work in progress for language teaching applications, requiring substantial teacher oversight to be effectively implemented in educational settings.

2.3. Detailed Training Analysis Results for GPT-4o

Note: This section provides detailed findings on GPT-4o performance.

The analysis of GPT-4o's performance demonstrated significant limitations requiring major teacher intervention. GPT-4 feedback showed very limited accurate output (indicated by green color in the original analysis) with frequent inaccuracies (red color) and numerous partial problems (blue color).

1. Training Results Summary

The training results for GPT-4o revealed considerable challenges across multiple evaluation dimensions. The model struggled with accurate error identification and classification, frequently provided direct corrections rather than guided feedback, and generated output at a readability level inappropriate for the target student population.

The most critical limitation was the low accuracy of feedback, with correct identifications (green-coded in the analysis) occurring far less frequently than partial problems (blue) or complete inaccuracies (red). This pattern remained consistent across repeated trials, suggesting systematic rather than random issues in GPT-4o's approach to language feedback.

Performance Metric	Results
Accurate feedback (green)	Limited accuracy, frequent errors
Partially correct feedback (blue)	Most common occurrence
Inaccurate feedback (red)	Frequent significant errors
Content feedback word count	Exceeds 200-word limit, close to 300 words
Readability level	Essay 1: B2-C1 (Flesch Reading Ease: 47.62), Essay 2: B2-C1 (Flesch Reading Ease: 47.13)
Consistency across trials	Low consistency, varying results in iterative trials with same prompt

2. Error Code Application and Feedback Approach

While GPT-4o exhibited numerous issues with error code application, these were more frequent than with other models tested. The most notable problems were in the areas of guided discovery and pedagogical approach, which was inconsistent in the model's output. Also, RW code was provided with other individual codes, which can cause a total confusion for the students.

The model inconsistently employed a guiding tone in its error descriptions, often providing direct corrections that would prevent students from engaging in self-correction. This approach contradicts best practices in language teaching, where metalinguistic feedback is preferred over direct correction. For example, instead of guiding students to discover errors themselves, GPT-4o would directly state "add 'an' before 'important'" - preventing students from developing their own error correction skills.

Content feedback consistently exceeded the appropriate length (well beyond the 200-word limit) while failing to effectively prioritize the most important areas for improvement. This lack of prioritization could lead to cognitive overload for students by presenting too many areas for correction simultaneously.

Some representative error examples include:

1. **Body 1, Line 1:** Directly stated missing article "an" before "important part" *Issue: Direct correction rather than guided discovery*
2. **Body 1, Line 2:** Used extra error codes for individual errors despite sentence requiring rewrite *Issue: RW should override all other error codes. This was consistently available across the error table.*
3. **Body 1, Line 3:** Directly provided correction "body's" for "bodies" *Issue: No guided discovery approach*

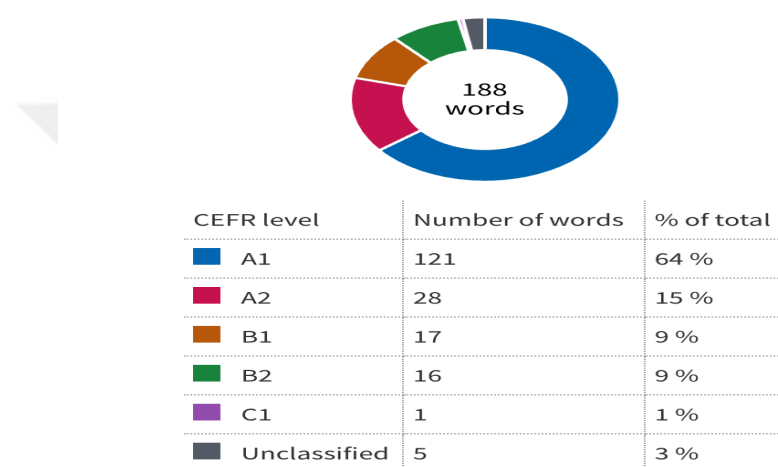
4. Readability Analysis of Content Feedback

Perhaps the most significant disadvantage of GPT-4o was its inability to generate feedback at an appropriate readability level for B1 students. Unlike Claude 3.5 Sonnet, which produced feedback at B1 levels, GPT-4o consistently generated feedback at B2-C1 levels of vocabulary and sentence complexity.

The readability analysis revealed an unbalanced distribution of vocabulary with considerable B1, B2, and even C1 vocabulary. Flesch Reading Ease scores of 47.62 and 47.13 for the two essays place the feedback in a range that is too complex for B1 level students, who would likely struggle with the language produced by the model.

This inappropriate readability level is a critical flaw for effective feedback, as overly complex explanations create barriers to student comprehension and diminish the instructional value of the feedback provided.

Essay 1: Oxford and Text Inspector Analysis Results for Readability



Summary ?		UPDATE SCORES
Sentence count	15 (Amend)	
Token count (excluding numbers)	194 (Amend)	
Type count (unique tokens, excluding numbers)	108 (Amend)	
Average sentence length	12.93 words	
Type/token ratio	0.56	
Number count	6 / 6 total digits	
Syllable count	335 (Amend)	
Words with more than 2 syllables	43	
Words with more than 2 syllables - Percentage	22.16	
Average syllables per sentence	22.33	
Average syllables per word	1.73	
Syllables per 100 words	172.68	
Readability Scores		
Flesch Reading Ease	47.62	
Flesch-Kincaid Grade	9.83	
Gunning Fog index	14.04	

Essay 2: Oxford and Text Inspector Analysis Results for Readability



CEFR level	Number of words	% of total
 A1	121	64 %
 A2	28	15 %
 B1	17	9 %
 B2	16	9 %
 C1	1	1 %
 Unclassified	5	3 %

Sentence count	14
Token count (excluding numbers)	247 (Amend)
Type count (unique tokens, excluding numbers)	137 (Amend)
Average sentence length	17.64 words
Type/token ratio	0.55
Number count	0 / 0 total digits
Syllable count	414 (Amend)
Words with more than 2 syllables	44
Words with more than 2 syllables - Percentage	17.81
Average syllables per sentence	29.57
Average syllables per word	1.68
Syllables per 100 words	167.61
Readability Scores	
Flesch Reading Ease	47.13
Flesch-Kincaid Grade	11.07
Gunning Fog index	14.18

Training Implications:

GPT-4o's performance suggests it should not be the preferred choice among the tested models for generating student feedback owing to significant need of intervention by

teachers, especially for the guiding tone and length of output for content feedback, compared to the minimal refinement needed for Claude 3.5 Sonnet outputs.

The key differences that make it less preferable include:

1. **Inappropriate Scaffolding:** The inconsistent tone and B2-C1-level readability fail to provide appropriate scaffolding for language learners, potentially overwhelming them.
2. **Feedback Delivery:** Technical inability to provide complete feedback in a single response worsens the user experience for both teachers and students.
3. **Readability of content feedback:** The readability of GPT-4o content feedback was overly complex and unsuitable for B1 level learners.

2.4. Researcher's Color-Coded Manual Evaluation of AI Feedback Accuracy (Claude 3.5 Sonnet, GPT-4o, and Gemini Advanced)

1.Results of Claude 3.5 Sonnet (a sample output):

Note: The feedback on the training papers was manually analyzed using a color-coded system designed by the researcher:

Green = correct coding,

Blue = overlooked/missed/partial codings

Red = wrong coding

This system helped evaluate Claude 3.5 Sonnet's error feedback accuracy.

Essay 1:

Essay Analyses, Common errors in the output and Overall evaluation

Essay 1

TITLE: Why should people have a balance**(SP)** diet?(C)

INTRODUCTION: Line 1: The importance of healthy nutrition has increased nowadays because to living a healthy life, a balanced diet is vital. RW Line 2: Balanced nutrition means getting the nutritions your body needs in the right amounts. RW Line 3: A balanced diet has many causes (WW) on the body. Line 4: Three of them reasons are mental health, improving digestion, lastly to live a healthy and long life. RW

BODY PARAGRAPH 1: Line 1: First of all, a balanced diet is A important part of maintaining health and living a long life. Line 2: A balanced diet provides the energy your body needs and helps you does(WF) your daily activities. Line 3: Carbohydrates and proteins are very important for the bodies(SP) energy production. Line 4: One taken in balanced way, it keeps the energy level high and helps us live a healthy and long life. RW

BODY PARAGRAPH 2: Line 1: Second of all, a balanced diet is very important for mental health. ✓
 Line 2: A Intake of nutrients such as omega-3 fats improves brain health and increases memory and focus abilities. Line 3: Additional or balanced diet can reduce stress and improve mood. RW

BODY PARAGRAPH 3: Line 1: FinallyP a balanced diet is very important for the healthy functioning of the organs in your body, especially the digestive system. Line 2: A healthy digestion affects everything. ✓ Line 3: The stomach and the intestines are the second brain. ✓

CONCLUSION: Line 1: In conclusionP there is no harm in a balanced diet.

ERROR TABLE

Error Table			
Location	Error Code(s)	Error Description	Sentence with Error or Correct Sentence
Title	SP, C	Is "balance" the right form to describe a diet? How do we usually capitalize titles in English?	Why should people have a balance**(SP)** diet?(C)
Intro, Line 1	RW	This sentence needs rewriting due to three or more errors or it lacks clarity. It has issues with grammar and word choice. How can you connect the importance of nutrition to the need for a balanced diet more clearly?	The importance of healthy nutrition has increased nowadays because to living a healthy life, a balanced diet is vital. RW
Intro, Line 2	RW	This sentence needs rewriting due to three or more errors or it lacks clarity. It has issues with word choice and spelling. What's the correct term for the things our body needs from food? How could you explain balanced nutrition more simply?	Balanced nutrition means getting the nutritions your body needs in the right amounts. RW
Intro, Line 3	WW	Is "causes" the right word here? What does a balanced diet do to the body?	A balanced diet has many causes(WW) on the body.
Intro, Line 4	RW	This sentence needs rewriting due to three or more errors or it lacks clarity. It has issues with grammar and structure. How could you list these three effects more clearly?	Three of them reasons are mental health, improving digestion, lastly to live a healthy and long life. RW
Body 1, Line 1	A	What word often comes before adjectives like "important"?	First of all, a balanced diet is A important part of maintaining health and living a long life.
Body 1, Line 2	WF	What form of the verb should follow "helps you"?	A balanced diet provides the energy your body needs and helps you does(WF) your daily activities.
Body 1, Line 3	SP	When we talk about energy production in general, do we use "body" or "bodies"?	Carbohydrates and proteins are very important for the

			bodies(SP) energy production.
Body 1, Line 4	RW	This sentence needs rewriting due to three or more errors or it lacks clarity. It has issues with grammar and clarity. What do you mean by "One taken in balanced way"? How can you explain the effects of a balanced diet more clearly?	One taken in balanced way, it keeps the energy level high and helps us live a healthy and long life. RW
Body 2, Line 1	✓	Well-written sentence. Good job!	Second of all, a balanced diet is very important for mental health. ✓
Body 2, Line 2	A	What word often comes before nouns like "Intake" in English?	A Intake of nutrients such as omega-3 fats improves brain health and increases memory and focus abilities.
Body 2, Line 3	RW	This sentence needs rewriting due to three or more errors or it lacks clarity. It has issues with clarity and grammar. What do you mean by "Additional or balanced diet"? How can you explain the connection between diet, stress, and mood more clearly?	Additional or balanced diet can reduce stress and improve mood. RW
Body 3, Line 1	P	What punctuation usually comes after words like "Finally" at the beginning of a sentence?	FinallyP a balanced diet is very important for the healthy functioning of the organs in your body, especially the digestive system.
Body 3, Line 2	✓	Well-written sentence. Good job!	A healthy digestion affects everything. ✓
Body 3, Line 3	✓	Well-written sentence. Good job!	The stomach and the intestines are the second brain. ✓
Conclusion	P	What punctuation usually comes after phrases like "In conclusion" at the beginning of a sentence?	In conclusionP there is no harm in a balanced diet.
Content Feedback (200 words): Strengths: Your essay has a clear structure with an introduction, three body paragraphs, and a conclusion. This organization helps readers follow your ideas easily. You've identified three main effects of a balanced diet: mental health, digestion, and longevity. For example, in your second paragraph, you mention omega-3 fats and their impact on brain health, which is a good specific detail. Areas of Improvement: - Thesis Statement: Can you make your main points clearer in the introduction? Try to list your three effects in one sentence. For instance, how could you combine the ideas of mental health, digestion, and longevity into one clear statement? - Supporting Details: In each body paragraph, try to add more specific information: - In paragraph 2, you mention omega-3 fats. Can you give an example of a food that contains omega-3 fats?			

- In paragraph 3, you say "A healthy digestion affects everything." Can you give an example of how good digestion helps our body?
 - How does a balanced diet help us live longer? Can you mention any health problems that a balanced diet might prevent?
3. Conclusion: Your conclusion is very brief. How could you summarize your main points here? Why do you think a balanced diet is important for everyone?

2.Results of GPT-4o (a sample output):

TITLE:

Why Should People Have a Balanced Diet?

INTRODUCTION:

Line 1: The importance of healthy nutrition has increased nowadays (**WW**) because to living a healthy life a balanced diet is vital (**WO**).**Line 2:** Balanced nutrition means getting the **nutritions** (**SP**) your body needs in the right amounts.**Line 3:** A balanced diet has many **causes** (**WW**) on the body (**WO**).**Line 4:** Three of **them reasons** (**WW**) are mental health, improving digestion, **lastly** (**TR**) to live a healthy and long life (**RW**).

BODY PARAGRAPH 1:

Line 1: First of all, a balanced diet is **important part** (**A**) of maintaining health and living a long life.
Line 2: A balanced diet provides the energy your body needs and helps you **does** (**WW**) your daily activities (**RW**).**Line 3:** Carbohydrates and proteins are very important for the **bodies** (**SP**) energy production.**Line 4:** **One taken** (**WW**) in a balanced way, it keeps the energy level high and helps us live a healthy and long life (**RW**).

BODY PARAGRAPH 2:

Line 1: Second of all, a balanced diet is very important for mental health. **Line 2:** Intake of nutrients such as omega-3 fats improves brain health and increases memory and focus abilities. **Line 3:** **Additional or balanced diet** (**WO**) can reduce stress and improve mood.

BODY PARAGRAPH 3:

Line 1: Finally, a balanced diet is very important for the healthy functioning of the organs in your body, especially the digestive system (**RW**). **Line 2:** A healthy digestion affects everything (**RW**).**Line 3:** **The stomach and the intestines are the second brain** (**IR**).

CONCLUSION: Line 1: In conclusion, there is no harm in a balanced diet (**RW**).

Error Table for Essay 1:

Section	Line Number	Error Code(s)	Error Description	Sentence with Error
			Feedback missing for the title	
INTRODUCTION	Line 1	WW, WO	"because to" is incorrect word usage; "to living" should be reordered.	The importance of healthy nutrition has increased nowadays because to living a healthy life a balanced diet is vital.
INTRODUCTION	Line 2	SP	"nutritions" should be "nutrients"	Balanced nutrition means getting the nutritions your

				body needs in the right amounts.
INTRODUCTION	Line 3	WW, WO	"causes" should be "effects"; incorrect word order.	A balanced diet has many causes on the body.
INTRODUCTION	Line 4	WW, TR, RW	"them reasons" should be "these reasons"; missing transition; sentence structure unclear, requires rewrite.	Three of them reasons are mental health improving digestion lastly to live a healthy and long life.
BODY PARAGRAPH 1	Line 1	A	Missing article "an" before "important part".	First of all, a balanced diet is important part of maintaining health and living a long life.
BODY PARAGRAPH 1	Line 2	WW, RW	"does" should be "do"; sentence requires rewrite due to unclear meaning.	A balanced diet provides the energy your body needs and helps you does your daily activities.
BODY PARAGRAPH 1	Line 3	SP	"bodies" should be "body's".	Carbohydrates and proteins are very important for the bodies energy production.
BODY PARAGRAPH 1	Line 4	WW, RW	"One taken" is unclear; requires rewrite.	One taken in balanced way it keeps the energy level high and helps us live a healthy and long life.
BODY PARAGRAPH 2	Line 3	WO	Incorrect word order; should be "A balanced diet or additional nutrients".	Additional or balanced diet can reduce stress and improve mood.
BODY PARAGRAPH 3	Line 1	RW	Sentence structure is unclear; requires rewrite.	Finally, a balanced diet is very important for the healthy functioning of the organs in your body, especially the digestive system.
BODY PARAGRAPH 3	Line 2	RW	Sentence structure is unclear; requires rewrite.	A healthy digestion affects everything.
BODY PARAGRAPH 3	Line 3	IR	Irrelevant information; does not support the paragraph's main idea.	The stomach and the intestines are the second brain.
CONCLUSION	Line 1	RW	Sentence structure is unclear; requires rewrite.	In conclusion, there is no harm in a balanced diet.

Content Feedback

Strengths of the Essay:

1. **Clear Topic:** The essay effectively introduces the topic of a balanced diet and its significance. The overall structure is logical, with an introduction, body paragraphs, and a conclusion.
2. **Relevant Points:** The essay touches on important aspects such as mental health, digestion, and the overall benefits of a balanced diet. These are relevant and valuable points to discuss in this context.

Areas of Improvement:

1. **Thesis Statement:** The introduction could benefit from a clearer thesis statement that concisely presents the main points of the essay. Consider revising the fourth sentence in the introduction to clearly outline the three main benefits of a balanced diet.
2. **Detailed Examples:** The body paragraphs would be stronger with more specific examples or explanations to support the main points. For instance, when discussing the importance of carbohydrates and proteins in Body Paragraph 1, you might mention specific foods that are rich in these nutrients and how they contribute to energy production.
3. **Conclusion Clarity:** The conclusion should summarize the main points more clearly. It would be helpful to restate the benefits of a balanced diet briefly before closing with a final thought.

3.Results of Gemini Advanced (a sample output):

Essay 1:

Error-Incorporated Essay

TITLE: The Importance of a Balanced Diet

INTRODUCTION: Line 1: Why should people have a **balance (SP)** diet? The importance of healthy nutrition has increased nowadays because **to (X)** living a healthy life, a balanced diet is vital. Line 2: Balanced nutrition means getting the **nutritions (S/P)** your body needs in the right amounts. Line 3: A balanced diet has many **causes (WW)** on the body. Three of **them (X)** reasons are mental health, improving digestion, lastly to live a healthy and long life.

BODY PARAGRAPH 1: Line 1: First of all, a balanced diet is **important part of (A)** maintaining health and living a long life. Line 2: A balanced diet provides the energy your body needs and helps you **does (S/V)** your daily activities. Line 3: Carbohydrates and proteins are very important for the **bodies' (P)** energy production. Line 4: **One (X)** taken in **balanced (A)** way, it keeps the energy level high and helps us live a healthy and long life.

BODY PARAGRAPH 2: Line 1: Second of all, a balanced diet is very important for mental health. Line 2: Intake of nutrients such as omega-3 fats improves brain health and increases memory and focus abilities. Line 3: **Additional (WW)** or balanced diet can reduce stress and improve mood.

BODY PARAGRAPH 3: Line 1: Finally a balanced diet is very important for the healthy functioning of the organs in your body, especially the digestive system. Line 2: A healthy digestion affects everything. The stomach and the intestines are the second brain.

CONCLUSION: Line 1: In conclusion there is no harm in a balanced diet.

Error Table

Location	Line Number	Error Code(s)	Error Description	Sentence with Error or Correct Sentence

TITLE	-	✓ (Correct)	Well-written title	The Importance of a Balanced Diet
INTRODUCTION:	1	SP (Spelling), X (Unnecessary Word)	You have misspelled a word, and there's an extra word that isn't needed. Can you spot them?	Why should people have a balance (SP) diet? The importance of healthy nutrition has increased nowadays because to (X) living a healthy life, a balanced diet is vital.
	2	S/P (Singular/Plural)	Double-check if this noun should be singular or plural.	Balanced nutrition means getting the nutritions (S/P) your body needs in the right amounts.
	3	WW (Wrong Word), X (Unnecessary Word)	You've used a word that doesn't quite fit the context, and there's an extra word that can be removed. Can you identify them?	A balanced diet has many causes (WW) on the body. Three of them (X) reasons are mental health, improving digestion, lastly to live a healthy and long life.
BODY PARAGRAPH 1	1	A (Article)	It seems like an article (a, an, the) might be missing or incorrect here. Check if you need one and which one to use!	First of all, a balanced diet is important part of (A) maintaining health and living a long life.
	2	S/V (Subject-Verb Agreement)	The subject and verb in this	A balanced diet provides the energy

			sentence don't agree. Make sure they match in number (singular or plural).	your body needs and you helps you does (S/V) your daily activities.
	3	P (Punctuation/Possessive Form)	It looks like there might be a punctuation or possessive form error here.	Carbohydrates and proteins are very important for the bodies' (P) energy production.
	4	X, A (Unnecessary Word, Article)	There's an unnecessary word here, and an article might be missing or incorrect. Can you find them?	One (X) taken in balanced (A) way, it keeps the energy level high and helps us live a healthy and long life.
BODY PARAGRAPH 2	1	✓ (Correct)	Well-written sentence	Second of all, a balanced diet is very important for mental health.
	2	✓ (Correct)	Well-written sentence	Intake of nutrients such as omega-3 fats improves brain health and increases memory and focus abilities.
	3	WW (Wrong Word)	The word you chose doesn't seem to fit the context. Try finding a more suitable word.	Additional (WW) or balanced diet can reduce stress and improve mood.
BODY PARAGRAPH 3	1	✓ (Correct)	Well-written sentence	Finally a balanced diet is very important for the healthy functioning of

				the organs in your body, especially the digestive system.
	2	✓ (Correct)	Well-written sentence	A healthy digestion affects everything. The stomach and the intestines are the second brain.
CONCLUSION:	1	✓ (Correct)	Well-written sentence	In conclusion there is no harm in a balanced diet.

Content Feedback

Strengths:

- **Clear Focus:** Your essay maintains a consistent focus on the positive effects of a balanced diet, which aligns well with the prompt.
- **Organized Structure:** The essay is well-organized, with each paragraph dedicated to a specific benefit of a balanced diet.
- **Concise Conclusion:** You wrap up your essay with a brief and impactful conclusion.

Areas for Improvement:

1. **Thesis Statement:** Your introduction could be strengthened with a clearer thesis statement that directly states the main benefits you will discuss in your essay. Consider revising the last sentence to something like, "A balanced diet offers numerous benefits, including improved physical and mental health, better digestion, and a longer life."
2. **Stronger Topic Sentences:** The topic sentences in each body paragraph could be refined to more clearly state the main point of each paragraph. For example, the topic sentence in the first body paragraph could be revised to something like, "A balanced diet provides the necessary energy for physical activity and overall well-being."
3. **More Supporting Details and Examples:** The body paragraphs could benefit from additional supporting details and examples to strengthen the arguments. Consider providing more specific examples of how a balanced diet improves mental health and digestion.
4. **Conclusion:** While your conclusion is concise, it could be further developed. Consider summarizing the key points of your essay and restating the overall importance of a balanced diet in a more impactful way.

APPENDIX-3: Sample Student Paper Employed for AI Feedback Generation

STUDENT ESSAY 1:

Every year the popularity of fast food restaurants is increasing. That is not a false statement, all day long individuals eats lots of fast food. Nowadays people try to avoid the fast foods in their lives, but this attempt finished with failure in the lots of times because fast food is every where in our century. There are several causes of increasing popularity of fast food restaurants, such as pop culture, adversiments and easy rotatable.

The first cause is "pop culture". In this culture, peopless always try act like a people who has a fame such as social media content productors, Hollywood star etc. And the fast foods restaurants give them a money to introduce their products for example they introduce product with recording video content and the induvadials who sees these videos. They try to eat like thems thanks to pop cultur.

The second cause is "adversiments" the fast food restaurants giving importance to adversiment because that has a positive impact to humans for example, there is no like McDonalds. Everybody knows this sentence because we hear this on McDonalds adversiments all the time. that reason why peoples eat lots of fast foods. Due to that the popularity of fast food restaurants is increased.

The 3rd and most important cause is easy to rotatable. Fast foods provides opportunity to people who is in rush easy to eat meals. So what does it's mean. That means, you can eat your meal in your car, at the park, in your house wherever you want. Also this meal doesn't come with fork and spoon. You need to use your hands to consume that. Finally you can order a meal in every way for example you can order with your phone with using apps. You can order in restaurant, you can open order in your car with using drive-thim service.

In introduction year by year popularity of fast foods restaurants increased because of pop culture,adversiment, and easy to rotatable.

ERROR-INCORPORATED ESSAY:

TITLE: (^)

INTRODUCTION: Line 1: Every year the popularity of fast food restaurants is increasing. (✓) Line 2: That is not a false statement, all day long **individuals eats (S/V)** lots of fast food. Line 3: Nowadays people try to avoid the fast foods in their lives, but this attempt **finished (T)** with failure in the **lots of (WW)** times because fast food is **every where (SP)** in our century. Line 4: There are several causes of increasing popularity of fast food restaurants, such as pop culture, **adversiments (SP)** and **easy rotatable (WW)**.

BODY PARAGRAPH 1: Line 5: The first cause is "pop culture". Line 6: **In (SP)** this culture, **peopless (SP)** always try **act (VF)** like **a (X)** people who **has (S/V)** a fame such as social media content **productors (SP)**, Hollywood **star (S/P)**etc. Line 7: And the fast foods restaurants **give (T)** them **a (X)** money to introduce their products **for example (P)** they introduce product with recording video content and the **induvadials (SP)** who sees these videos. Line 8: They try to eat like **thems (SP)** thanks to pop **cultur (SP)**.

BODY PARAGRAPH 2: Line 9: The second cause is "**adversiments (SP)**" **the (X)** fast food restaurants giving importance to **adversiment (SP)** because that has a positive impact to humans **for example (P)**, there is no like McDonalds. Line 10: Everybody knows this sentence because we hear this on McDonalds **adversiments (SP)** all the time. Line 11: **that (C)** reason why **peoples (S/P)** eat lots of fast foods. Line 12: Due to that the popularity of fast food restaurants **is (T)** increased.

BODY PARAGRAPH 3: Line 13: The **3rd (WF)** and most important cause is **easy to rotatable (WW)**. Line 14: Fast foods **provides (S/V)** opportunity to people who **is (S/V)** in rush easy to eat meals. Line 15: So what **does (WW)** it's mean. Line 16: That means, you can eat your meal in your car, at the park, in your house **wherever (WW)** you want. Line 17: Also this meal doesn't come with fork and spoon. (✓) Line 18: You need to use your hands to consume that. (✓) Line 19: Finally you can order a meal in every way **for example (P)** you can order with your phone with using apps. Line 20: You can order in restaurant, you can **open (WW)** order in your car with using **drive-thim (SP)** service.

CONCLUSION: Line 21: **In (C)** introduction year by year popularity of fast foods restaurants increased because of pop culture, **adversiment (SP)**, and **easy to rotatable (WW)**.

ERROR TABLE GENERATED BY AI

Location of Error	Error Code(s)	Error Description	Sentence with Error
TITLE	(^)	Consider adding a title to your essay. How could you summarize your main topic in a few words?	N/A
INTRODUCTION - LINE 2	(S/V)	Check if "eats" agrees with "individuals". How do we usually form verbs with plural subjects?	That is not a false statement, all day long individuals eats (S/V) lots of fast food.
INTRODUCTION - LINE 3	(T, WW, SP)	Think about the tense of "finished". Is it describing a current situation? Consider a different phrase for "lots of". Is "every where" one or two words?	Nowadays people try to avoid the fast foods in their lives, but this attempt finished (T) with failure in the lots of (WW) times because fast food is every where (SP) in our century.
INTRODUCTION - LINE 4	(SP, WW)	Check the spelling of "adversiments". Is "easy rotatable" the phrase you want to use? Think about how we describe food that's easy to eat quickly.	There are several causes of increasing popularity of fast food restaurants, such as pop culture, adversiments (SP) and easy rotatable (WW) .
BODY 1 - LINE 6	(SP, SP, VF, X, S/V, SP, S/P)	Is "In" capitalized correctly? Check the spelling of "peopless". Should "act" have "to" before it? Do we need "a" before "people"? Does "has" agree with "people"? Check the spelling of "producers". Should "star" be plural?	In (SP) this culture, peopless (SP) always try act (VF) like a (X) people who has (S/V) a fame such as social media content producers (SP) , Hollywood star (S/P) etc.
BODY 1 - LINE 7	(T, X, P, SP)	Is "give" in the right tense? Do we need "a" before "money"? Consider punctuation after "for example". Check the spelling of "induvadials".	And the fast foods restaurants give (T) them a (X) money to introduce their products for example (P) they introduce product with recording video content and the induvadials (SP) who sees these videos.
BODY 1 - LINE 8	(SP, SP)	Check the spelling of "thems" and "cultur". How do we usually form plural pronouns?	They try to eat like thems (SP) thanks to pop cultur (SP) .
BODY 2 - LINE 9	(SP, X, SP, P)	Check the spelling of "adversiments". Do we need	The second cause is "adversiments (SP)" the

		"the" before "fast food"? Check the spelling of "adversiment". Consider punctuation after "for example".	(X) fast food restaurants giving importance to adversiment (SP) because that has a positive impact to humans for example (P) , there is no like McDonalds.
BODY 2 - LINE 10	(SP)	Check the spelling of "adversiments". How do we usually spell this word?	Everybody knows this sentence because we hear this on McDonalds adversiments (SP) all the time.
BODY 2 - LINE 11	(C, S/P)	How do we usually start sentences? Is "peoples" the correct plural form of "people"?	that (C) reason why peoples (S/P) eat lots of fast foods.
BODY 2 - LINE 12	(T)	Is "is" the right tense here? Are you talking about a past or present increase?	Due to that the popularity of fast food restaurants is (T) increased.
BODY 3 - LINE 13	(WF, WW)	How do we usually write "3rd" in formal writing? Is "easy to rotatable" the phrase you want? Think about how we describe food that's easy to eat quickly.	The 3rd (WF) and most important cause is easy to rotatable (WW) .
BODY 3 - LINE 14	(S/V, S/V)	Does "provides" agree with "Fast foods"? Does "is" agree with "people"?	Fast foods provides (S/V) opportunity to people who is (S/V) in rush easy to eat meals.
BODY 3 - LINE 15	(WW)	Is "does" the word you want to use here? How do we usually ask what something means?	So what does (WW) it's mean.
BODY 3 - LINE 16	(WW)	Is "wherever" the word you want? Consider if you need a different word to express your idea.	That means, you can eat your meal in your car, at the park, in your house wherever (WW) you want.
BODY 3 - LINE 19	(P)	Consider adding punctuation after "for example". How do we usually separate examples in a sentence?	Finally you can order a meal in every way for example (P) you can order with your phone with using apps.
BODY 3 - LINE 20	(WW, SP)	Is "open" the word you want to use with "order"? Check the spelling of "drive-thim". What's the correct term for ordering food from your car?	You can order in restaurant, you can open (WW) order in your car with using drive-thim (SP) service.

CONCLUSION LINE 21	- (C, SP, WW)	How do we usually start sentences? Check the spelling of "adversiment". Is "easy to rotatable" the phrase you want? Think about how we describe food that's easy to eat quickly.	In (C) introduction year by year popularity of fast foods restaurants increased because of pop culture, adversiment (SP) , and easy to rotatable (WW) .

CONTENT FEEDBACK:

Strengths: Your essay has a clear structure with an introduction, three body paragraphs, and a conclusion. You've identified three main causes for the increasing popularity of fast food restaurants: pop culture, advertisements, and convenience. This organization helps readers follow your main ideas.

Areas for Improvement:

1. **Thesis Statement:** Try to make your main points clearer in the introduction. For example, you could write: "The increasing popularity of fast food restaurants is due to three main factors: influence of pop culture, effective advertising, and convenience for consumers."
2. **Supporting Details:** In each body paragraph, consider adding more specific examples:
 - In the pop culture paragraph, can you mention a specific celebrity or social media influencer who has promoted fast food?
 - In the advertisement paragraph, can you describe a memorable fast food advertisement campaign?
 - In the convenience paragraph, try to explain more clearly how fast food fits into people's busy lifestyles.
3. **Language Use:** Pay attention to spelling, especially for repeated words like "advertisement". Also, try to use more varied vocabulary. For example, instead of saying "fast food" repeatedly, you could use phrases like "quick meals" or "convenient dining options".

APPENDIX-4: Expert Opinions

Table 1. Expert Panel Experts

No	Title	Area of Expertise	Date
1	Assoc. Prof. Dr.	ELT (Assessment and Evaluation)	30.08.2024
2	Lecturer	English	31.08.2024
3	Lecturer	English	01.09.2024
4	Asst. Prof. Dr.	ELT	01.09.2024
5	Dr.	Language Consultancy and ICT	05.09.2024
6	Lecturer	English Language Teaching / Assessment & Evaluation / PD	05.09.2024

Table 2. Evaluation of AI Tools

Değerlendirme Alanı	Uzman Değerlendirmesi	Öne Çıkan Öneriler
1. AI Prompt (Yapay Zekâ Komutu)	5 Uzman: Uygun	- LW yerine TR kullanımının değerlendirilmesi
	1 Uzman: Düzeltilmeli (Uzman 5)	- "Possible modifications" başlığı eklenmesi
		- Alt başlıkları kolay tanımlanabilir şekilde kalın yazı kullanılması
2. Hata Kodları Dosyası	6 Uzman: Uygun	-
3. Hata Kodları Detaylı Açıklamalar	6 Uzman: Uygun	-
4. Öğretmen Geribildirim Örnekleri	6 Uzman: Uygun	-
5. Yapay Zekâ Eğitiminde Öğrenci Yazıları	6 Uzman: Uygun	-

Table 3. Evaluation of Training and Support Materials

Değerlendirme Alanı	Uzman Değerlendirmesi	Öne Çıkan Öneriler
6. Öğretmen Eğitim Programı	5 Uzman: Uygun	- Program öncesi öğretmen bakış açılarını anlamak için
	1 Uzman: Düzeltilmeli (Uzman 5)	test/görüşme yapılması
		- "Mom's test" uygulaması önerisi
7. Word Dokümanı Şablonu ve Talimatlar	6 Uzman: Uygun	-
8. Google Excel Tablosu	6 Uzman: Uygun	-
9. Dereceli Puanlama Yönergesi	6 Uzman: Uygun	-

Table 4. Evaluation of Interview Questions

Değerlendirme Alanı	Uzman Değerlendirmesi	Öne Çıkan Öneriler
10.1. Genel Bilgi Soruları	6 Uzman: Uygun	-
10.2. Doğruluk ve Kullanılabilirlik Soruları	5 Uzman: Uygun	-
	1 Uzman: Düzeltilmeli (Uzman 5)	
10.3. Geribildirim İnceleme Süreci Soruları	6 Uzman: Uygun	-
10.4. Öğretmen Geliştirme Süreci Soruları	5 Uzman: Uygun	-
	1 Uzman: Düzeltilmeli (Uzman 5)	
10.5. Avantaj/Dezavantaj Soruları	6 Uzman: Uygun	-
10.6. Sınıf-İçi Kullanım Soruları	5 Uzman: Uygun	- Senaryo içeren soruların eklenmesi önerisi
	1 Uzman: Düzeltilmeli (Uzman 5)	
10.7. İyileştirme Önerileri Soruları	6 Uzman: Uygun	-
10.8. Görüşme Kapanışı Soruları	6 Uzman: Uygun	-

Table 5. Detailed Suggestions from Expert 5 (Language Consultancy and ICT)

Alan	Öneriler
AI Prompt (Yapay Zekâ Komutu)	1. For labelling errors, I suggest using LW instead of TR since it might have different connotations for the learner and might be misleading.
	2. I recommend that Claude.ai should also include different, more complex structures and vocabulary versions of how some phrases can be told based on IELTS or any other proficiency exam prompted which could enable the learner to improve his/her complexity and diversity in writing skills. This section might be titled as 'Possible modifications'.
	3. If subtitles are bolded, it can be easier to diagnose the errors. Similar colours without using any italic or bold letters are quite not user-friendly.
Öğretmen Eğitim Programı	Bence programa başlamadan önce öğretmenlerin bakış açılarını anlamak, geçmiş tecrübelerini daha iyi gözlemlemek için bir test veya interview yapılmalı. Mom's test okumanızı öneririm çünkü aksi takdirde RQ3 kısmındaki öğretmenlerin karşılaştıkları zorluklar vs noktasında veya eğitim programının faydalı olması noktasında ön yargılı sorularla karşılaşılabilir. Öğretmenler iyi görünmek için cevaplarda duyulmak istenen cevapları da verebileceğinden konudan bağımsız bir gözlem ve inceleme yapılmalı ve geçmiş tecrübeleri sorulmalıdır.
Sınıf-İçi Soruları	Bu kısımda cevaplama zor veya istenilen cevabı verme ihtimali olan bir kısım. Bence senaryo içeren sorular entegre edilebilir buraya.

Table 6. Positive Feedback from Expert 6 (Testing Department)

Alan	Geri Bildirimler
AI Prompt (Yapay Zekâ Komutu)	Hata kodu ve içeriğe yönelik geri bildirim komutları detaylı, beklentileri karşılar nitelikte sunulmuştur.
Hata Kodları Dosyası	Hata Kodları, B1 seviyesi öğrencilerinin yazma performanslarında düzeltmeleri beklenen hatalar öngörülmüş ve kapsamlı bir şekilde kodlar oluşturulmuştur.
Hata Kodları Detaylı Açıklamalar	Örnekler, gerçekçi, açıklama dili detaylı ve uygundur.
Öğretmen Geribildirim Örnekleri	Geribildirim tablosu amaca uygun, ayrıntılı bir şekilde hazırlanmıştır. Hatanın tanımının yapılması, bu hata için yönlendirmede bulunulması ve hatalı bölümün tekrar tabloda gösterilmesi, açık ve yönlendirme konusunda başarılı olmuştur.
Yapay Zekâ Eğitiminde Öğrenci Yazıları	Başarılı bir uygulama olduğu düşünüyorum.
Öğretmen Eğitim Programı	Detaylı ve amaca etkili bir şekilde hizmet edecek bir program oluşturulmuş.

Word Dokümanı Şablonu ve Talimatlar	Belgeler ve talimatlar açık anlaşılır ve amaca yönelik hazırlanmıştır.
Google Excel Tablosu	Raporlamalar ve yönergeler açık, anlaşılır ve detaylıdır.
Dereceli Puanlama Yönergesi	Amaçlar, açıklamalar ve yönlendirmeler başarılı bir şekilde hazırlanmıştır.
Görüşme Soruları	Sorular amaca yönelik etkili bir şekilde tasarlanmıştır.



APPENDIX-5: Teacher Training Materials

Note: This appendix contains training materials, including the training programme, handouts, review guidelines, and sample teacher reviews used during the piloting phase.

1.Overview of Training Sessions

The Overview of Four-Day Training Session

Day 1: Introduction to AI Feedback and AI Feedback Evaluation and Enhancement Process

Session Duration: 1.30 hours

Materials: PPT slides, General AI Feedback Review and Enhancement Guide for Teachers, color-coded review guide for teachers, sample teacher review of error and content feedback, and structured review table.

1. PPT Presentation (15 minutes): Overview of the research aims and AI-generated feedback, including benefits and limitations from the literature.
2. Introduction to General Teacher Guide for Enhancing and Reviewing AI Feedback for Accuracy & Explanation of the Process (40 minutes):
 - Introducing the general guide outlining the two phases: enhancing AI feedback in Word (error-incorporated paper & content feedback) and documenting enhancements in the structured review table (Handout-1).
 - Introduction to Word format for reviewing and enhancing AI feedback (Handout-2: color-coded review format and instructions).
 - Demonstration of Sample AI Feedback and teacher enhancement (Handout-3: sample teacher refinement).
3. Explanation of Structured Review Table Content (35 minutes):
 - Overview of the Structured Review Table referring to steps in Handout-1 (10 minutes).
 - Walkthrough of Spreadsheet headings and their functions (10 minutes).
 - Explanation of dropdowns: acceptance levels, rationale options, and their link to the rubric (Handout-4).

Day 2: Reviewing and Refining Sample AI Feedback Using Review Table

Session Duration: 1 hour

Material: Sample AI feedback as an assignment.

Teachers begin with the color-coded review of the error-incorporated paper and content feedback in Word, while simultaneously documenting changes in the structured review table. The table also includes detailed error descriptions that are not reviewed in the Word document but should be reviewed in the table.

Day 3: Discussion of Reviews Completed by Teachers and Peer Review

Session Duration: 30 minutes

Activity: Peer review and calibration (30 minutes)

- Teachers present sample reviews from pilot study.
- Group discussion to analyze pilot reviews and make calibration decisions (15 minutes).

Day 4: Using Rubric to Evaluate AI Feedback and Peer Review for Calibration

Session Duration: 1.15 hours

Introduction (20 minutes):

- Explanation of rubric criteria and performance levels (Handout-5: rubric evaluation instructions).
- Note: Rubric explanations are also available in a voiced PPT for later reference.

Workshop (55 minutes):

- Activity: Rubric-based evaluation of previously reviewed AI feedback (15 minutes).
- Small group discussion of ratings and rationales (10 minutes).
- Group-wide calibration check (10 minutes).
- Q&A and completion of training form (10 minutes).

Training Schedule for Teacher Reviews and Data Collection

Week	Activity	Description
Week 1	AI Feedback Generation by the researcher	AI Feedback Generation from Claude
Week 1-2	Training & Initial Standardization	Through training session and initial standardization meeting (pilot study with 1 paper to familiarize with review process and initial standardization for rubric evaluation before starting the review)
Week 2	Initial Standardization & Calibration Session 1	Evaluators completed assignment of reviewing AI feedback by using color-coded and structured review table, and rate sample AI feedback. So, this session will be done to discuss ratings and align understanding for review of AI feedback and rubric evaluation.
Week 2-4	Paper Review (1-5)	Teachers review first 5 papers
Week 3	Calibration session 2	Calibration session for clarifying doubts and remaining consistent in analyses.
Week 5	Brief Standardization & Mid-point Calibration 3	Mid-point standardization meeting and calibration session to realign evaluators if needed
Week 6-8	Paper Review (6-10)	Teachers review papers 6 through 10

Week 8	Brief meeting for any lingering questions	Brief meeting for any lingering questions
Week 9-10	Completion of reviews	Completion of reviews
Week 10-12	Individual Interviews	Conduct individual interviews with teachers

2. PowerPoint Presentation Slides Summary

Slide 1: Brief Information on Thesis Topic and Research Objectives

Thesis Topic: This thesis aims to explore how AI-generated feedback can be integrated into the feedback process used by teachers. Specifically, it seeks to evaluate how accurate and useful AI feedback is and how teachers filter and refine this feedback, and the pros and cons of this process.

Research Questions:

1. How do teachers refine AI-generated feedback with their own expertise and insights?
2. How do teachers evaluate the accuracy of AI-generated feedback for both errors and content in student essays?
3. What are the advantages and challenges teachers experienced in refining AI-generated feedback and their perspectives on its future classroom use?

Answering these research questions will complement one another, offering a comprehensive understanding of the research phenomenon:

- Enhancing AI Feedback: Exploring how teachers modify AI feedback and identifying common practices to blend AI-generated feedback with their insights.
- Teachers' Perceptions on AI Feedback Accuracy: Exploring teachers' insights on the accuracy of AI-generated feedback for correcting errors and content in essays.
- Challenges and Advantages of Enhancing AI Feedback: Exploring the challenges teachers face and the benefits they perceive when enhancing AI-generated feedback.

Slide 2: Purpose of this training

- Purpose: You are participating in this training to test how AI feedback can be used to assist teachers for feedback provision.
- Objective: By reviewing and improving the feedback provided by AI, you will help determine whether the AI tool can effectively support the feedback process.

Slide 3: The Rise of AI in Provision of Feedback

Slide 4: What are your biggest concerns when providing feedback as an instructor?

Slide 5: Concerns

- Time
- Consistency in Quality Standards
- Large Number of Students

Slide 6: The Rise of AI in Providing Feedback

Content: While traditional sources like teacher and peer feedback are valuable, they face limitations in scalability, timely feedback delivery, and consistency. In contrast, AI feedback offers a promising alternative by providing timely, scalable, and consistent feedback, addressing these challenges.

Slide 7: Limitations of AI Feedback

Content: AI is not free from biases, inconsistencies, and ethical concerns. The accuracy of AI feedback remains a significant issue, as it can make mistakes. Therefore, a teacher Filter is necessary, as human expertise is still required.

Slide 8: AI Feedback vs. Teacher Completion

Table Format:

Limitations of AI Feedback: Biases, Inconsistencies, Ethical Concerns, Lack of Nuanced Contextual Understanding, Lack of Emotional and Motivational Support, Trust Issues.

Teacher Completion: Expert Knowledge, Ethical and Privacy Oversight, Nuanced Contextual Understanding, Emotional and Motivational Support, Trust and Credibility.

Slide 9: Training Snapshot

Training Overview: The training is designed to teach you the procedure for improving AI feedback and using the review table for this purpose. It will familiarize you with the tools, AI feedback, and help you evaluate your review table findings more reliably.

Slide 10: Tasks

You have 3 tasks:

1. Color-coded review: Improve the paper with errors and content feedback as you would in your daily practice. Track the time.
2. Structured review table: Complete feedback review documentation.
3. Rubric Evaluation.

3. Training Materials

Handout 1: Teacher Review Handbook

Task 1: Color-coded Review for Teacher Refinements

Color Coding System for reviewing error codes:

Review error codes, check overlooked errors and also error code placement for their accuracy. Use color coding below:

- Incorrect codes or wrong error code placements (in red)
- Adding missing/overlooked codes (in blue), including tick, RW, or other individual codes.
- Identify and mark unnecessary codes for removal (highlight in yellow, add “x” next to the part you want to remove)
- Perform complete sentence rewrites (place RW code at the end, in blue)
- Execute partial rewrites (underline part to be rewritten, place RW code at the end of it, in blue)

Reviewing Content Feedback

Review content feedback for technical and pedagogical accuracy (relevance, constructiveness, guiding tone for student improvement):

- Add your comment for content to explain why you accept, partially accept, or reject the feedback.
- Clearly state where AI feedback needs improvement and add your suggestions to improve them. Also comment on the points you liked in the review.

Task 2: Structured Review and Documentation

AI Feedback Accuracy: Document your edits and changes to the Google Sheet table based on your Word document review.

Indicate Acceptance Level: Select Accept / Partially Accept / Reject from the drop-down list.

Consider both technical and pedagogical correctness: When indicating accuracy, do not just consider 'correctness.' Consider guidance, constructiveness, and clear/actionable suggestions.

Provide Rationale for Change(s):

- Choosing multiple rationale options is okay. Prioritize 'Reject (1)' if feedback has significant flaws. Otherwise, select 'Partially Accept (2).'
- Add 'Other Rationale(s)' if needed.

Reviewing Content Feedback:

- Review content feedback for technical and pedagogical accuracy (relevance, constructiveness, guiding tone for improvement).
- Add your comment to the Google review form to explain why you accept, partially accept, or reject the feedback.
- Clearly state where AI feedback needs improvement (do this in the Google review form to avoid duplicating effort).

Shortcut Definitions for Constructive and Guiding Tone in Error Descriptions

Constructive Tone:

Phrases like “*consider*,” “*do you think*,” “*try*,” and “*think*” soften the tone of feedback, making it constructive (i.e., positive, non-critical, and not overly directive). This encourages students to self-correct.

Commands such as “*check it*” or “*rewrite it*,” or using a directive tone, do not reflect a constructive tone. It's essential to soften the language and prompt students to reflect and make corrections on their own.

Guiding Tone:

A guiding tone involves not giving away the correct answer directly. Instead, it frames questions, hints, or prompts that support student self-correction.

Improved Feedback (Rating 3)

"How do we usually start sentences? Look at the word "adversiment". Does this spelling seem correct? Is "easy to rotatable" the phrase you want? Think about how we describe food that's easy to eat quickly."

- ✓ All guiding questions/prompts
- ✓ No direct instructions

Original Feedback (Rating 2)

"How do we usually start sentences? Check the spelling of "adversiment". Is "easy to rotatable" the phrase you want? Think about how we describe food that's easy to eat quickly."

- ✓ 3 guiding questions/prompts
- ✗ 1 direct instruction

Key Difference

The direct instruction "Check the spelling" has been replaced with guiding questions:

✗ "Check the spelling of "adversiment"."



✓ "Look at the word "adversiment". Does this spelling seem correct?"

Selecting Acceptance level in Structured Review table

Low Accuracy – Reject (Rating: 1)

Mark the review as 'Reject' if ANY of the following conditions are present, as they require significant teacher intervention:

1. Completely incorrect error codes or redundant feedback despite some correct parts.
2. Major overlooked errors and codes, especially when combined with minor ones.
3. Multiple overlooked errors in one sentence.
4. Completely misplaced error codes that mislead students.
5. Demotivating or non-constructive tone.
6. Answer-giving tone that prevents student self-correction.

Mid Accuracy – Partially Accepted (Rating: 2)

Feedback is partly correct but includes minor flaws or lacks clarity:

- Mixed Accuracy: Combination of correct and incorrect feedback elements.
- Minor overlooked errors or elements.
- Mixed or incorrect error code order/placement.
- Vague or non-specific error explanations.
- Incomplete/inconsistent guidance.
- Constructive tone needs some improvement.

High Accuracy – Accepted (Rating: 3)

Feedback is correct, clear, and needs minimal changes:

- Complete and accurate identification with no significant errors overlooked.
- Guiding tone used, not giving direct answers.
- Constructive, encouraging, and actionable suggestions.
- Only minor rephrasing or clarification may be needed.

Task 3. Rubric Evaluation Guidelines

A. Default Scenarios

Google Review Pattern → Rating

- Mostly 'Reject' → 1
- Mostly 'Partially Accept' → 2
- Mostly 'Accept' → 3

B. Mixed Scenarios

Scenario → Rating → Notes

- Majority Accept + few Partial/Reject → 3 → Lean toward Accept
- Even mix of Accept, Partial, and Reject → 2 → Lean toward Partial
- Mostly Partial + some Accept + few Reject → 2 → Lean toward Partial
- Partial + Accept > Reject → 2 → Lean toward Partial
- Majority Reject + few Partial/Accept → 1 → Lean toward Reject
- Reject > Partial + Accept combined → 1 → Lean toward Reject

Annotated Examples for Rubric Application

Example 1: Rating 3 (Accurate Feedback)

Student's sentence: "Balanced diet has numerous benefit for mind and body."

AI Feedback: "Great job in your thesis statement ✓ Do you think 'benefit' is the correct form here? Consider using the plural form of benefit."

Annotation:

- ✓ Correct identification of the error (plural noun needed)
- ✓ Positive reinforcement at the beginning
- ✓ Clear explanation of the issue
- ✓ Guiding tone and no direct correction, prompting student to find the answer

Example 2: Rating 2 (Partially Accurate Feedback)

Student's sentence: "The balanced diet has many affects on people's mind and body. It contributed to mental wellbeing and improved physical health. However, it also improves your physical appearance."

AI Feedback: "Your paragraph provides different reasons for choosing balanced diet. There are a couple of small errors to address:

1. 'affects' should be 'effects' - remember, 'effect' is usually a noun.
2. 'contributed' should be 'contribute' - this is the present tense form. You might want to write more on each point with specific examples to strengthen your idea."

Annotation:

- ✓ Correctly identifies two grammar errors
- ⚠ Directly provides corrections instead of guiding the student
- ✓ Offers a good suggestion for improvement (expanding with examples)

Example 3: Rating 1 (Incorrect Feedback)

Student's sentence: "A balanced diet is essential for maintaining good health. It provides the body with necessary nutrients and helps prevent various diseases. Moreover, a balanced diet can improve mental wellbeing and increase energy levels."

AI Feedback: "Your paragraph needs significant improvement. Here are the corrections:

1. Change 'is essential' to 'are essential' because 'diet' is plural.
2. Replace 'provides' with 'provide' to match the plural subject.
3. 'Moreover' is not the right transition word here. Use 'In addition' instead.
4. The last sentence is completely wrong and should be removed."

Annotation:

- ✗ Incorrectly identifies errors in grammatically correct sentences
- ✗ Provides misleading corrections that would introduce errors
- ✗ Uses an overly critical tone without any positive reinforcement
- ✗ Suggests removing a valid and relevant sentence
- ✗ Fails to recognize and reinforce the student's correct use of language and content
- ✗ Offers direct corrections instead of guiding the student to self-correction.

Training (Piloting Feedback Form)

Dear Teachers,

Please take a few minutes to complete this Google Form to share your thoughts on the clarity, usefulness, and overall effectiveness of the materials—Part 1: color-coded review and Google Sheet review table, and Part 2: rubric—as well as your evaluation of the training process.

Your feedback is invaluable in refining these tools to better support you in evaluating AI-generated feedback.

Thank you once again for your valuable time and input!

Regards,

Instructor Burcu Gündoğdu

1. Name & surname
2. Your position in Antalya Bilim University

Color-coded review Instructions:

- How clear and useful were the color-coded review format and instructions?
- Do you have any suggestions for improvement (e.g., color coding, clarity of checklist, clarity of instructions)?

Rubric Feedback:

- How clear, usable, specific, and non-redundant are the criteria in the rubric?

- Were the criteria and descriptors in the rubric clear and easy to understand? If not, please specify the parts you found unclear.

Structured Review Table:

- How easy was it to use the Google Sheet review table for AI feedback evaluation?
- Any suggestions for improving teacher review in Google?

Usability of the Rubric:

- Was the rubric easy to use for assessing AI-generated feedback on student papers?
- What challenges did you face, and what aids would help clarify the ratings or improve your interpretation? (Challenges may include understanding and differentiating ratings, interpreting descriptors, or dealing with mixed acceptance level performance.)

Relevance and Efficiency:

- Were the rubric criteria relevant for evaluating AI feedback quality?
- Do you have any suggestions for additional criteria or descriptors?
- Were there any criteria you found redundant? If so, please specify.

Overall Evaluation of Training:

- What did you think of the overall training content and activities?
- Were they relevant and effective in helping you understand how to evaluate AI feedback?

Thank you for your feedback!

4. A Sample from Piloting Task: Color-coded teacher refinement

*Note: Teacher employed color-coding for their refinements: **blue** for overlooked errors, **red** for incorrect codes, and **yellow** for redundant codes.*

Color-coded Review

ERROR-INCORPORATED ESSAY:

TITLE: (^)

INTRODUCTION: Line 1: Fast food restaurants **being** (VF) so popular nowadays. Line 2: **There is** (S/V) fast food companies in the world and it (**Pr**) is increasing every day. Line 3: Most people **likes** (S/V) consume (VF) fast foods. Line 4: As a result of this, it has some effects on people. (✓) Line 5: **There is** (S/V) three effects of increasing popularity of fast food restaurants; increasing health problems^{**}; **decreasing consume** (WF) of traditional foods and being more lazy.

BODY PARAGRAPH 1: Line 6: The first effect of increasing popularity of fast food restaurants is increasing health problems. (✓) Line 7: Fast food **increasing** (VF) possibility of being **obesity** (WF) and the other diseases such as **diabet** (SP) and heart diseases. Line 8: Fast foods include so much sugar, salt and harmful oils (**S/P**). (✓) Line 9: Because of this, **It** (C) **affect** (S/V) on (**X**) our metabolism. Line 10: Our metabolism **being** (VF) slower (**P**) so we can **gaining** (VF) weight easily. Line 11: For example^{**}, **fast food companies is** (S/V) so popular in (**^**) U.S.A and (**TR**) they struggle with obesity.

BODY PARAGRAPH 2: Line 12: The second effect of increasing popularity of fast food restaurants is **dicreasing** (SP) **consume** (WF) of traditional foods. Line 13VD: Most people **prefers** (S/V) eating fast food in a day because working **take** (S/V) a lot of time. Line 14: Cooking traditional foods (**^**) **more** (WFx) difficult than **order** (VF) a (**X**) food from fast food restaurants. Line 15: **Although** (TR) we have knowledge about includes (WF) in traditional foods. (**RW**) Line 16: Maybe we should try **consume** (VF) more traditional foods because **it** (S/V) is healthier than fast foods.

BODY PARAGRAPH 3: Line 17: The third effect of increasing popularity of fast food restaurants is being more lazy. (✓) Line 18: Maybe fast food is a (**A**) opportunity for using more time (**RW**) but most people consume fast food because **of** (**X**) it is easy (**^**). Line 19: **Therefor** (SP) we called being lazy about this behavior. (**RW**) Line 20: Since increasing fast food restaurants, (**RW**) we tend (**VF**) to be lazy, but we should stop it (**X**) (**^**) and **doing** (VF) some foods for **ourselves** (SP).

CONCLUSION: Line 21: **In** (SP) conclusion, **increasing** (WFx) (**^**) of fast food restaurants **have** (S/V) many effects on us. Line 22: We should be careful about how much eating (VF) fast food and **trying** (VF) to decrease (**^**) eating (**X**) fast food (**RW**). Line 23: Fast food **lead** (S/V) to many diseases but if we consume less **than** (WWx) it will be (**x**) not (^) a problem.

5. A Sample from Piloting Task: Structured Review Table for (error-coded feedback)

<i>Section& location</i>	<i>Error Code</i>	<i>Error description by AI</i>	<i>Sentence with error code</i>	AI feedback (Accepted/ Not Accepted/Partially Accepted)	Teacher Edits/Changes/Removals	rationale for teacher changes
TITLE	(^)	Consider adding a title to your essay. How could you summarize your main topic in a few words?	N/A	accepted (no change needed)		no 15
INTRODUCTION - LINE 1	(VF)	Think about the verb form of "being". What tense do we usually use for general statements?	Fast food restaurants being (VF) so popular nowadays.	accepted (no change needed)		no 15
INTRODUCTION - LINE 2	(S/V)	Check if "There is" agrees with "companies". How do we form this phrase with plural nouns?	There is (S/V) fast food companies in the world and it is increasing every day.	partially accepted (some changes needed)	There is (S/V) fast food companies in the world and it (PR) is increasing every day. It is not clear what "it " refers to. pronoun error.	no 8, no 11
INTRODUCTION - LINE 3	(S/V)	Does "likes" agree with "Most people"? Consider the usual form for plural subjects.	Most people likes (S/V) consume fast foods.	partially accepted (some changes needed)	Line 3: Most people likes (S/V) consume (VF) fast foods. There is no feedback about the correct form of the verb "consume"	no 8, no 11
missing line 4	no tick code			partially accepted (some changes needed)	missed the line. no tick code provided in excel.	no 11

INTRODUCTION - LINE 5	(S/V, WF)	Check if "There is" agrees with "three effects". Is "consume" the correct form here? Consider using the noun form.	There is (S/V) three effects of increasing popularity of fast food restaurants; increasing health problems; decreasing consume (WF) of traditional foods and being more lazy.	accepted (no change needed)		no 15
missing line 6	no tick code			partially accepted (some changes needed)	missed the line. no tick code provided in excel.	no 11
BODY 1 - LINE 7	(VF, WF, SP)	Think about the verb form of "increasing". Is "obesity" the correct form here? Check the spelling of "diabet".	Fast food increasing (VF) possibility of being obesity (WF) and the other diseases such as diabet (SP) and heart diseases.	accepted (no change needed)		no 15
missing line 8	no tick code			partially accepted (some changes needed)	missed the line. no tick code provided in excel.	no 11
BODY 1 - LINE 9	(C, S/V)	How do we usually capitalize "it" in sentences? Does "affect" agree with "it"?	Because of this, It (C) affect (S/V) on our metabolism.	partially accepted (some changes needed)	Overlooked code "x" for "on". On should be removed. Here is no clarification about the difference between "affect sth" and "have an effect on sth"	no 8, no 11

BODY 1 - LINE 10	(VF, VF)	Consider the verb forms of "being" and "gaining". What tense do we use for general statements?	Our metabolism being (VF) slower so we can gaining (VF) weight easily.	rejected (major changes needed)	Our metabolism being (WW) slower (P) so we can gaining (VF) weight easily. Student should use "getting slower" instead of "being slower" and there must be a comma before so.	no 1, no 8
BODY 1 - LINE 11	(S/V)	Does "is" agree with "companies"? How do we form verbs with plural subjects?	For example, fast food companies is (S/V) so popular in U.S.A and they struggle with obesity.	partially accepted (some changes needed)	"AND" should be changed with ""SO" Student should use "the" before USA	no 8, no 11
BODY 2 - LINE 12	(SP, WF)	Check the spelling of "decreasing". Is "consume" the correct form here? Consider using the noun form.	The second effect of increasing popularity of fast food restaurants is decreasing (SP) consume (WF) of traditional foods.	accepted (no change needed)		no 15
BODY 2 - LINE 13	(S/V, S/V)	Does "prefers" agree with "Most people"? Does "take" agree with "working"?	Most people prefers (S/V) eating fast food in a day because working take (S/V) a lot of time.	accepted (no change needed)		no 15
BODY 2 - LINE 14	(WF, VF, X)	Is "more" the correct form here? Think about the form of "order" after "than". Do we need "a" before "food"?	Cooking traditional foods more (WF) difficult than order (VF) a (X) food from fast food restaurants.	rejected (major changes needed)	AI didn't give feedback about the missing verb and there is also incorrect feedback about "more"	no 2, no 11

BODY 2 - LINE 15	(TR, WF)	Is "Although" the transition you want to use here? Is "includes" the correct form?	Although (TR) we have knowledge about includes (WF) in traditional foods.	rejected (major changes needed)	"Although (TR) we have knowledge about includes (WF) in traditional foods." It is not a full and meaningful sentence, so it should be rewritten fully, but RW code overlooked and no description provided for this major problem.	no 2, no 4
BODY 2 - LINE 16	(VF, S/V)	Think about the form of "consume" after "try". Does "it" agree with "foods"?	Maybe we should try consume (VF) more traditional foods because it (S/V) is healthier than fast foods.	rejected (major changes needed)		no 15
missing line 17	no tick code			partially accepted (some changes needed)	missed the line. no tick code provided in excel.	no 11
BODY 3 - LINE 18	(A, X)	Which article do we use before words starting with a vowel sound? Is "of" necessary before "it is easy"?	Maybe fast food is a (A) opportunity for using more time but most people consume fast food because of (X) it is easy.	rejected (major changes needed)	Two major overlooked mistakes regarding word choice overlooked and this can affect understanding of sentence if not corrected. It should be " using time efectively" and "it is easy to order"	no 2, no 4
BODY 3 - LINE 19	(SP)	Check the spelling of "Therefor". How do we usually spell this transition word?	Therefor (SP) we called being lazy about this behavior.	rejected (major changes needed)	The sentence is incorrect and needs full rewriting and yet code RW overlooked here. No error description	no 2, no 11

					provided for this major problem.	
BODY 3 - LINE 20	(VF, SP)	Think about the form of "doing" after "should". Check the spelling of "ourselves". How do we usually form this reflexive pronoun?	Since increasing fast food restaurants, we tend to be lazy, but we should stop it and doing (VF) some foods for ourselves (SP).	rejected (major changes needed)	overlooked codes, wrong code (VF) which should be (WW), and overlooked RW code. The correct sentence should be " Since the increase of fast food restaurants, we have tended to be lazy, but we should stop eating fast food and start cooking for ourselves."	no 2, no 4, no 11
CONCLUSION - LINE 21	(SP, WF, S/V)	Check the spelling of "In". Is "increasing" the correct form here? Does "have" agree with "increasing"?	In (SP) conclusion, increasing (WF) of fast food restaurants have (S/V) many effects on us.	partially accepted (some changes needed)	"increase in the number of restaurants" should be the correct form. I think if missing word " number" added, the meaning will be more clear.	no 8, no 11
CONCLUSION - LINE 22	(VF, VF)	Consider the forms of "eating" and "trying" in this context. Do we need the -ing form here?	We should be careful about how much eating (VF)fast food and trying (VF) to decrease eating fast food.	rejected (major changes needed)	The sentence needs rewriting. The correct sentence should be "We should be careful about the amount of fast food we eat and trying (VF) to decrease the amount of fast food"	no 2, no 4
CONCLUSION - LINE 23	(S/V, WW)	Does "lead" agree with "Fast food"? Is "than" the word you want to use here?	Fast food lead (S/V) to many diseases but if we consume less than (WW) it will be not a problem.	partially accepted (some changes needed)	The correct sentence should be " If we consume less, it will not be a problem."	no 8

6.A Sample from Piloting Task: Structured Review Table for (content-feedback)

<i>Content feedback (within 200 words)</i>	<i>Strenghts</i>	<i>Areas needing improvement</i>	<i>AI feedback (Accepted/ Not Accepted/Partially Accepted)</i>	<i>Teacher Edits/Changes/Removals- Write where you will edit (strenghts, areas needing improvement, or both, and write your change(s) if there is any.</i>	<i>Choose the number from drop-down list for your rationale for Teacher Changes</i>	<i>Additional Rationale (If "Other")</i>
	Strengths: Your essay has a clear structure with an introduction, three body paragraphs, and a conclusion. You've identified three main effects of the increasing popularity of fast food restaurants: health problems, decrease in traditional food consumption,	Areas for Improvement: 1. Thesis Statement: Try to make your main points clearer in the introduction. For example, you could write: "The increasing popularity of fast food restaurants has three main effects: it leads to health problems, decreases consumption of traditional foods, and promotes a more sedentary lifestyle." 2. Supporting Details: In each body paragraph, consider adding more specific examples: o In the health problems paragraph, can you mention specific nutrients that are often lacking in fast food? o In the traditional foods paragraph, can you give an example of a traditional dish that people might be eating less	accepted (no change needed)	There are some direct feedback for the thesis statement and language use.	no 23	

	and increased laziness. This organization helps readers follow your main ideas.	often? o In the laziness paragraph, try to explain more clearly how fast food contributes to a sedentary lifestyle. 3. Language Use: Pay attention to subject-verb agreement and verb forms. For example, in your introduction, "There is fast food companies" should be "There are fast food companies." Also, try to use more varied vocabulary. Instead of repeating "fast food" so often, you could use phrases like "quick meals" or "convenience foods."				

APPENDIX-6: Color-coded review format and guidelines for teachers

Instructions for Color-Coded Teacher Review

In the color-coded review, you'll review the AI-generated feedback using color-code scheme, focusing on the error codes. Your task is to correct any inaccuracies and refine the feedback for clarity and effectiveness. Ensure that correct sentences are properly marked, and that error codes are accurately identified and placed. Mark any changes in red to distinguish them from the original AI feedback.

Format for the color-coded review is composed of raw AI error-incorporated text and a second page where you need to perform your color-coded review:

PAGE 1: AI-Generated Original Feedback Word Document

Title:

Paper No: [Student's ID or Paper Number]

AI-Generated Original Feedback

Instructions:

This document contains the original feedback generated by AI. Use it as a reference when making refinements. The error codes and feedback should remain as they are in this document for review before making any changes.

[AI ORIGINAL FEEDBACK INSERTED HERE BY THE RESEARCHER]

PAGE 2: Teacher Refinements to AI Feedback Word Document

Title:

Paper No: [Student's ID or Paper Number]

Teacher Refinements to AI Feedback

Instructions:

Instructions:

This document is where you should insert your refinements to AI-generated feedback using the color-coding system for error feedback. For content feedback, write your comments in the structured review table (google sheets) to avoid duplicating your work.

Note: Perform both review simultaneously to ensure you do not forget your reviews for each piece of feedback.

Color Coding Scheme for Reviewing error-coded feedback:

- ☐ incorrect codes or wrong error code placements (in ● red)
- ☐ Adding missing/ overlooked codes (in ● blue), including tick, rw, or other individual codes.
- ☐ Identify and mark unnecessary codes for removal (highlight in yellow, add "✕" next to the part you want to remove)
- ☐ Perform complete sentence rewrites (place RW code at the end, (in ● blue)
- ☐ Execute partial rewrites (underline part to be rewritten, place RW code at the end of it, (in ● blue)

Reviewing content-related feedback:

This review will be conducted using a structured review table in the form of a Google Sheet.

You need to review content feedback for technical and pedagogical accuracy of feedback (relevance+constructiveness+ guiding tone for student improvement).

- ☐ Add your comment to structured review table for content feedback to explain why you accept or partially accept or reject the feedback.
- ☐ State where AI feedback needs improvement clearly, and write strengths of feedback as well as your suggested changes directly into the form).

[TEACHER REFINEMENTS INSERTED HERE BASED ON INSTRUCTIONS]

Timeline and Goal:

1. Aim to analyze 1-2 papers per week. During midterm and final exam weeks, no calibration meetings or paper analyses will be conducted.

2. If the researcher identifies a problem in reviews, follow-up check sessions can be conducted individually, in addition to group sessions for calibration.

For any questions or clarifications, please contact me in person or send your question via e-mail [burcu.gundogdu@antalya.edu.tr].

APPENDIX-7: Structured Review Table

Instructions for Reviewing AI Feedback Accuracy with Structured Review Table

Dear Participants,

In the Structured Review Table (Google Sheets provided), you will mainly do three things to systematically document your refinements:

1. Record any changes you made to the AI feedback in the Word review under the section "Teacher Edits/Changes/Removals."
2. Evaluate the AI feedback by selecting one of the three acceptance levels.
3. Provide a rationale for your changes using the codebook, which is integrated into the dropdown list of numbered rationales. If your reason is not listed, please don't forget to write your own rationale.

Acceptance levels are as follows:

- **Reject (1) / Low-Accuracy:** feedback that is mostly incorrect, unhelpful, unguiding, unconstructive that requires significant teacher intervention.
- **Partially Accept (2) / Mid-Accuracy:** feedback that is somewhat helpful but still needs some teacher intervention to correct inaccuracies, clarify vague points, or mixed tone in guidance or constructiveness.
- **Accept (3) / High-Accuracy:** feedback that effectively identifies errors, provides clear guidance, and is constructive, requiring minimal or no changes from the teacher.

For minimal changes that qualify for an **Accept (3)** rating, consider the following examples:

1. *Adding a brief clarification to an already correct description.*
2. *Slightly rephrasing for better clarity or removing an information unnecessary without changing the meaning.*

3. *Adding a small motivational comment to reinforce the constructive tone.*

4. *Providing an additional example to support the existing feedback.*

However, if more substantial changes are needed, you should **consider using** *Partially Accept (2)* instead.

Handling Multiple Rationales for a Single Piece of Feedback

If you select multiple rationales for a single piece of feedback, and if it contains any of the following significant flaws given below, prioritize a "Rejected (1)" rating:

LOW- ACCURACY (rejected-rating 1)

Major errors requiring major teacher intervention, as impacts the feedback's fundamental purpose.

1. **Wrong Error Code/Description:** Incorrect or misleading redundantly given error code or description.
2. **Major overlooked error:** A sentence with global problem is provided local errors rather than “rewriting” code to fix the sentences’ meaning, structure or coherence problem. The individual codes are not helping fix the problem (sentence fragment, run-on, missing main verbs or subjects etc.). Besides that, the code of multiple overlooked errors can also overwhelm students, and this is also classified under major error.

e.g. Missing feedback for lack of main verb, subject, or not given RW code can lead to global problems untreatable with individual codes.
3. **Wrong Content Feedback/ description:** Incorrect or irrelevant content-related feedback or description.
4. **Entirely wrong error code placement:** misplaced error code.
5. **Lack of Guiding:** Fails to guide or support improvement for self-correction, as it provides exact corrected version or solution.
6. **Lack of Constructive Tone:** come across as overly directive or critical, demotivating student for self-correction.

MID-ACCURACY (Partially Accepted- rating 2)

Correct in parts but includes minor flaws or lacks full clarity:

1. **Mixed Accuracy:** The error code / content feedback involves mix of correct and incorrect elements.
2. **Minor overlooked errors:** missing or overlooked elements in the error identification, accompanied by correctly identified ones.
3. **Mixed error code order, or correct and wrong error code placement:** both correct and incorrect code placement or mixed order of error codes.
4. **Vague or non-specific feedback:** not very clear feedback on how to correct the error (vague), where the error is not clear (non-specific).
5. **Incomplete/ inconsistent Guidance:** Correctly identifies errors but is only partially guiding or lacks detail. Or it partially provides guidance while provide exact answers/ solutions.
6. **Constructive Tone Needs Improvement:** Feedback is supportive but not fully encouraging or actionable.

When to Consider it Partial: If feedback has mix of correct/incorrect error identification, or if it includes vague, non-specific, mixed of correct or wrong error code placement or inconsistent guidance or a constructive tone needing teacher “partial” enhancement.

HIGH ACCURACY (Accepted- rating 3)

Feedback is fully correct, actionable, and motivating, needing minimal teacher modifications:

1. **Complete accuracy:** Correct error/content identification and description. No errors/ strengths / areas needing improvement are overlooked.
2. **Guiding tone: Correctly** describe the error/ content issue without giving away exact corrected version or solutions.
3. **Constructive Tone:** Encourages students to improve with specific actionable suggestions.

4. **Minimal Changes Required** Only slight rephrasing, additions for clarifications, or small motivational comments required.

Key Differences: Unlike partial feedback, “Accept (3)” is fully actionable, requires minimal changes, and fully supports student learning.

Final Note: Before starting your review in the Google Sheets, please refer to the **General Guide's "Task 2" section** to ensure you correctly carry out the Structured review table documentation steps. Use the explanations provided under "**Selecting the Correct Acceptance Level**" as a reference throughout your evaluation of both error and content feedback. **Consulting Examples:** During your review, you may also refer to examples provided for Error and Content Feedback criteria, which illustrate feedback at low, mid, and high accuracy levels. These examples will help guide your evaluation and provide clarity on what each accuracy level looks like.

Numbered Rationale Options: This table includes numbered rationale options assigned for easier selection in the Excel dropdown, and also categorized as two table: error and content feedback.

Table 1: Error-coded Feedback

Reject (1) – Low Accuracy	Partially Accept (2) – Mid Accuracy	Accept (3) – High Accuracy
1. Incorrect error code: The error code is incorrect. Wrong or redundantly given code, which is wrong for that word or sentence, fall under this category.	7. Mix of correct and incorrect error identification: Some errors in the sentence or a word are identified correctly while some are identified incorrectly	15. Accurate and actionable error feedback: The feedback is correct, clear, guiding, constructive, and actionable, either requiring no changes or only minor adjustments as described in “minimal changes” section.
2. The presence of major or multiple overlooked errors: When AI assigns multiple local codes, but this is not fixing sentence-level problem. This happens when AI fails to address global issues affecting sentence structure, meaning and coherence and cohesion, and does not assign RW (rewriting) code.	8. Minor Overlooked error(s): Some surface-level errors (i.e. SP, X, P) in the sentence-level or word-level are overlooked, while others are successfully identified. This requires teachers to complete/ add overlooked minor kind of errors to student paper.	-----

	9. Mixed/confusing error code order: Error codes identified are listed in a confusing order in “error codes” section in error table, leading to some confusion for evaluators	-----
3. Wrong error code placement: The error code(s) is placed incorrectly in the sentence, making it difficult for the student to understand what the error is and how to correct it	10. Mixed of correct and wrong error code placement: Several errors have been identified in one sentence, but some error codes are placed after or before of that word that need clarification (e.g., a mix of correct and incorrect placements)	
4. Entirely inaccurate error description: The error code and its description is incorrect, irrelevant, or misleading, making it difficult for students to act on it.	11. Partially- accurate error description: The AI feedback has missing, incomplete and irrelevant error description for the error. Besides that, mismatch can occur between AI feedback and structured review table, owing to missing elements. -Mismatch between AI feedback in word document and Structured Review (google sheets) table for error identification. e.g. Case 1: AI feedback and error detection in the structured review table match, but error description is not complete or include irrelevant error description. e.g. Case 2: The ticked sentence in the word document shows that the sentence was acknowledged as correct, which is positive and complete feedback. However, the error description missed providing feedback (e.g., confirming the sentence is correct and praising the student briefly). However, this is not misleading the student, as there is nothing for student correct in that sentence.	-----
. -----	12. Vague/non-specific feedback: The error description is vague or confusing, making it difficult for the student to understand what needs to be corrected or improved (e.g.,	-----

	"check form of the word" without specifying which form or section of the essay is not specified, - which body etc.)	
5. Lack of constructive tone: Feedback only identifies errors without any constructive tone to help student to improve or correct it. Instead, it comes across more as too directive, like 'do it,' 'rewrite it', 'check it' etc. Case 1: e.g. Ticked sentence was not provided in Case 2: e.g. Only identifies errors in an overtly directive tone.	13. Partially constructive tone in error description: While the feedback correctly identifies mistakes and attempts to guide the student toward improvement, the tone could still be more constructive (i.e., it can ask more positively or soften the language to encourage self-correction). For example, with a partially constructive tone: "you need to use a different word here for clarity." or "This point is not clear. You might need to rephrase it." (This is not overly directive or critical but not very encouraging either, as it does not include "consider", "try", phrases inviting to think, and also does not specify the aspect to fix.)	-----
6. Lack of guiding tone/ direct corrections in error description: Case 1: The error description provides the exact correct version without giving the student a chance to correct it themselves. Case 2: AI missed the line and did not give any error description although the sentence has errors to fix.	14. Mix of guidance and direct corrections: The feedback sometimes guides the student on how to improve but other times provides exact corrections. This can happen in error descriptions where one error includes guiding, while the other has direct solution.	-----

Table 2: Content Feedback

Reject (1) – Low Accuracy	Partially Accept (2) – Mid Accuracy	Accept (3) – High Accuracy
16. Entirely inaccurate content feedback: The content feedback is incorrect, irrelevant or misleading, making it difficult for students to address the content issues.	19. Partially accurate content feedback: The feedback addresses some relevant and useful points, but misses others or includes some irrelevant points.	23. Accurate and actionable content feedback: The feedback is correct, clear, guiding, constructive, and actionable, either requiring no changes at all or only minor adjustments as described in “minimal changes” section.
-----	20. Vague/ nonspecific feedback: The content feedback is vague or confusing, making it difficult for the student to understand what needs to be corrected or improved. (E.g. “the thesis statement is unclear” which does not explain how to improve or does not specify which section or part in the essay.)	-----
17. Lack of constructive tone: The feedback is overly critical, focusing only on errors, not mentioning strengths. It comes across more as critical, too directive, like 'do it,' “your thesis statement is wrong”, “check it” etc.	21. Inconsistent constructive tone: The feedback identifies areas needing improvement but doesn't acknowledge some strengths. The tone could still be more constructive (i.e., less directive, less critical, and more encouraging for self-correction). For example, “This point is not clear. You might need to rephrase it.” (This is not overly directive or critical but not very encouraging either.)"	-----
18. Lack of guiding tone/ direct corrections: The content feedback provides the exact correct version without giving the student a chance to correct it themselves (i.e. rewriting the sentence directly.)	22. Mix of guidance and direct corrections: The feedback sometimes guides the student on how to improve but also provides exact corrections.	-----

APPENDIX-8. Rubric for Evaluating Accuracy of AI-generated Feedback

Rubric For Evaluating Accuracy of AI-Generated Feedback

Aim: This rubric is designed to evaluate the accuracy of AI-generated feedback for student essays. It focuses on not only the correctness of error identification but also pedagogical correctness: constructive and guiding tone, aligning with teacher-like accuracy and style. Thus, the rubric also aims to determine the extent of teacher intervention needed to refine the AI feedback.

Instructions for use:

- **Consult the term explanations** if you are unsure about the concepts (e.g. what is guiding feedback?) and review the **Pre-rubric Guidance** before applying **Rubric-in- Use** section for evaluating AI-generated feedback.
- If still uncertain about your rating, **refer to the Structured Review Tables (google sheets)** for your **overall acceptance level** (Accept, Partially Accept, or Reject), and use default or mixed scenario rating for taking informed decisions.

→ **Default Scenarios** were designed to help you make informed decisions more easily by guiding you through the acceptance results in your structured review table.

- If most entries are marked "**Reject**" → **Rate 1**
- If most entries are marked "**Partially Accept**" → **Rate 2**
- If most entries are marked "**Accept**" → **Rate 3**

→ **Mixed Scenarios:** In cases where there is no clear majority among "Reject," "Partially Accept," or "Accept" in the structured review table, consider the following:

Scenario	Rating	Notes
Majority Accept + a few Partial, / a few Rejects	3	Lean toward Accept
Even mix of Accept, Partial, and Reject	2	Lean toward Partial
Mix of mostly partial + some accept items, and less amount of reject items	2	Lean toward Partial
Partial + Accept countings together outnumber Reject countings: Rating 2 – Lean toward Partial	2	Lean toward Partial
Majority Reject + a few Partial+ a few accept	1	Lean toward Reject
Reject countings outnumber Partial countings, even with Accept countings included	1	Lean toward Reject

Explanations for the Terms:

1. Guiding & Constructive Tone:

- **Guiding Tone:** Feedback that helps students discover correct answers on their own, using questions or hints (try, consider, how do you think, can you... etc.) rather than directly providing “exact” answers.
- **Constructive Tone:** Positive and supportive feedback that highlights strengths, marks correct answers and encourages improvement by focusing on 1-2 strengths and 2-3 areas for growth.

2. Clear and actionable suggestions:

- **Clear Suggestions:** Specific, easy-to-understand feedback that is not vague and avoid broad suggestions. For example, instead of saying, “This sentence is unclear,” it explains the problem and indicate what to change or how to change: “clarify your main point and add an example to paragraph 1, or for error description, instead of saying, “consider using the correct form of it”, it can suggest correct form of it, e.g. adjective form of the word.
- **Actionable Suggestions:** Practical feedback that provides specific steps for improvement. Instead of just noting generally the problem issue, such as “your paragraphs lacks transitions words and add transition words here”, it should offer alternatives like, “To improve coherence, try using transition words like ‘however’ or ‘therefore’ to link your ideas more effectively.

Pre-Rubric Guidance: Read Before Applying Ratings!!!		
Rating	Description	Guiding explanations
1 (Low accuracy)	Incorrect Feedback: The AI feedback is entirely incorrect or misleading, requiring major corrections.	<i>Use this rating if feedback contains “major” significant errors or flaws, including:</i> • Incorrect error identification, redundant given codes, wrong descriptions, or wrong placement of error codes. • Major overlooked errors at global level (sentence-meaning level), and hence codes individually not helping the fix the sentence. Or use of multiple overlooked errors in a piece of feedback at local level which discourages student to correct their mistake (i.e. not providing RW (rewrite) in such cases lead to multiple coding overwhelming for students to interpret the feedback). Note: Global errors are important codes teachers consider fixing the sentence. These are: not providing code for missing main verb, subject, or both, not providing RW (rewriting) code for full rewriting although the problem is in meaning-level. Not providing rw code, but providing individual codes, which not actually allow for correcting the meaning-related problem in the sentence. • Inaccurate content feedback; mislabel strong or relevant elements in the writing as weak or irrelevant. • Lack of guiding tone (direct corrections) • Lack of constructive tone (too directive, critical, and not positive/ supporting suggestions for self-correction/ improvement)
2 (Mid-accuracy)	Partially Accurate Feedback: The AI feedback has a mixture of correct and incorrect elements, requiring partial changes to be useful.	<i>Use this rating if the feedback needs some significant changes to be fully effective, including:</i> • Error identification involves mix of correct and incorrect elements. • Minor overlooked errors at local level, more of mechanics like sp, s/p etc. Partial RW at phrase or word levels are not global and hence can be considered as minor error. • Error description is mix of relevant/ irrelevant/incomplete/ missing/ vague or nonspecific points. • Error codes can involve wrong error code order or mix of correct/ wrong error code placement. • Partially correct content descriptions, mix of relevant/ irrelevant/incomplete or missing points.

		• Inconsistent guiding tone, mix of guiding for self-correction and direct exact corrections. • Inconsistent constructive tone, mix of supportive and too directive/ overly critical comments.
3 (High accuracy)	Accurate Feedback: The AI feedback is mostly accurate and relevant, with minimal/no changes needed.	<i>Use this rating if the AI feedback is mostly accurate and usable with minimal changes. It may only need minor adjustments, such as:</i> • Brief clarification to an already correct description. • Slight rephrasing for clarity or removing a phrase or sentence without changing the meaning. • Adding a small motivational comment to reinforce a constructive tone. • Providing an additional example to support the feedback.

Rubric- in use: Apply this rubric below for rating AI feedback's accuracy			
Rating	Feedback Description	Explanation	Guidance for Teachers
1(Low accuracy)	Incorrect Feedback The feedback is entirely incorrect or misleading, requiring major corrections.	Rating 1: Incorrect Feedback (Low- accuracy) Error Identification: ☐ Entirely incorrect code in identifying errors, or redundant codes unnecessarily placed for correct elements. ☐ Major overlooked errors that are not identified, thereby preventing student corrections. It can include global errors go unnoticed other than error affecting structure and coherence. Additionally, this can involve multiple errors remain unidentified. Error Placement: ☐ Many errors are placed incorrectly (e.g., completely misplaced error code). Error Description: ☐ Descriptions of errors are entirely incorrect, making it difficult for students to correct the errors. Content Evaluation: ☐ Inaccurate evaluation of content strengths and weaknesses. ☐ Suggestions are entirely incorrect and misleading, not allowing students to address the content issues. Tone for error descriptions and content: ☐ Lacks a guiding tone; provides direct corrections instead of helping students discover the correct answers on their own. ☐ Overly critical/too directive unconstructive feedback that is not supportive of self-correction or improving due to unsupportive, critical and too directive tone. It mostly points out mistakes or areas for improvement but rarely, if ever, mentions strengths.	Guidance: Not all criteria may be relevant in every case. If several criteria are met, the feedback should be rated as "Incorrect." Extent of teacher intervention/filter: ☐ Major revisions by the teacher are necessary. The feedback significantly misguides the student or requires substantial teacher correction.

2(Mid-accuracy)	Partially Accurate Feedback The AI feedback has a mixture of correct and incorrect elements, requiring partial changes to be useful.	Rating 2: Partially Accurate Feedback (Mid-accuracy) Error Identification: ☐ Mix of correct and incorrect error identification ☐ Some overlooked minor errors on surface-level errors, easy to detect by teacher intervention. Error code order or placement problem: ☐ Some errors are placed incorrectly or need clarification (e.g. mix of correct and incorrect placements, or error code(s) identified in mixed order). Error Description: ☐ Descriptions of errors are partially accurate, or they may have incomplete unclear/vague, non-specific, or irrelevant aspects, making it less actionable. Content Evaluation: ☐ Some content strengths and weaknesses are accurately described, but some points are missing or overlooked. ☐ Suggestions for improvement may have incomplete/ unclear/vague, non-specific, or irrelevant aspects, making it difficult for students to address the content issues. Tone for error descriptions and content: ☐ Inconsistent guidance, with a mix of guiding and direct corrections by providing exact answer. ☐ Inconsistent constructive feedback is less effective because it requires a more supportive tone to make the feedback positive and encouraging for self-correction. It may identify errors and areas for improvement but doesn't consistently highlight strengths (e.g., overlooking correct sentences that should be marked with a tick or failing to mention certain strengths).	Guidance: Not all criteria may be relevant in every case. Extent of Teacher Intervention/Filter: ☐ Partial teacher revisions required. If feedback is mixed, with both correct and incorrect elements, and requires moderate teacher intervention, it should be rated as "Partially Accurate."
3(high accuracy)	Accurate Feedback The AI feedback is mostly accurate and relevant, with minimal/no	Rating 3: Accurate Feedback (High accuracy) Error Identification: ☐ Errors are correctly identified with no overlooked errors. Error code Placement: ☐ Most errors are placed correctly, with only minor exceptions that don't affect understanding (e.g., error codes placed directly on the verb rather than between subject and verb).	Guidance: Not all criteria may be relevant in every case. Extent of Teacher Intervention/Filter: ☐ No or minimal adjustments needed by the teacher. If the feedback is generally accurate, requires little to no correction,

	changes needed.	Error Identification: Descriptions of errors are highly accurate as they are mostly clear, specific and relevant, allowing students to correct easily. Content Evaluation: ☐ Accurate evaluation of content strengths and weaknesses. ☐ Clear, relevant and actionable suggestions for improvement, enabling students to address content issue effectively. Tone for Error Descriptions and Content: ☐ Consistently guiding without directly providing corrections. ☐ Balanced and constructive feedback consistently provides positive and encouraging suggestions for improvement and self-correction. In content, it recognizes strengths while identifying errors or areas for improvement (e.g., marking correct sentences with a tick and presenting strengths).	and provides useful guidance, it should be rated as "Accurate."
Comment on this:	For any rating, specify the area(s) where AI feedback excels or needs improvement to justify the rating you have given:		

APPENDIX-9: Interview Questions

Interview Questions

General Questions for Background Information

1. Could you pls tell me a bit about yourself and your current position in the university?
2. How long have you been teaching, and what are the typical sizes of your classes?
3. How familiar are you with Generative AI platforms (e.g., GPT, Claude, or Gemini)? Have you used these tools for feedback on students' writing essays, or other writing-related feedback activities?

Interview Questions on the Evaluation of the AI Feedback System, pros and cons of refining it, and suggestions for usable AI-generated feedback

The following interview questions were designed to understand teachers' approaches to integrating AI feedback with their own insights, their perceptions of its accuracy and pros and cons of enhancing AI feedback, perspectives on its future use and suggestions for improving AI feedback.

(Focus: Teachers' experiences with AI feedback, its accuracy, pros/cons, future use, and suggestions for improvement.)

1. AI Feedback Accuracy

Think about the papers you reviewed in terms of accuracy,

1. What types of corrections did you make most often? Were there recurring errors or patterns you noticed?
2. Can you describe a specific content-related change you made to AI feedback?

What was incorrect or unclear, and how did you fix it?

3. Can you recall a time when AI feedback was particularly accurate? What made that feedback accurate and helpful?

2. Pros and Cons of Enhancing AI Feedback

Think about your experience refining AI feedback:

4. What was difficult about AI feedback and refining it?

5. What aspects of the refinement process made it easier or more helpful?

3. Classroom Application and Its Future Use

6. After trying this out, would you use AI feedback in your future teaching? What factors would make you more or less likely to use it?

7. What support would you need to use AI feedback effectively in your classroom? (e.g., additional training, technical support, or collaboration with colleagues)

4. Suggestions for Improvement and Closing Thoughts

8. Looking back at your experience with AI feedback, what specific improvements would make it more accurate and better aligned with how teachers give feedback in the classroom?

9. Is there anything else you'd like to share about your experience with the AI feedback system?

Thank you for your time and insights. Your feedback is invaluable in helping me to understand and improve the AI feedback system for teaching purposes.

APPENDIX-10: Sample analyses from a Teacher Reviews (color-coded and structured review samples)

***Note:** This appendix involves a sample (T6, paper 3) color-coding for refinement of AI-generated feedback (1): blue for overlooked errors, red for incorrect codes, and yellow for redundant codes, and a (2) sample structured review table Sample completed by the same teacher (T6, paper 3), and (3) Structured Review (Content Feedback).*

1. Color-coded Review

ERROR-INCORPORATED ESSAY:

TITLE: (^)

INTRODUCTION: Line 1: In the world, (^) fast food restaurants **are increase (VF)** day by day. Line 2: Fast food restaurants **are (X)** don't give up (^), and they always find new food types. (✓) Line 3: Therefore, people prefer these restaurants. (✓) Line 4: Also, fast food restaurants **are (X)** started in the U.S.A **the (X) first time (RW)**. Line 5: **Why (WO) these restaurants most (X) popular?** Line 6: People prefer these restaurants because they **tell (WW)** three causes for this topic: These causes are accessing easily, providing job opportunities and preparing easily (RW).

BODY PARAGRAPH 1: Line 7: The first **most (WW)** cause **people (WW)** is people can easily **the (X)** fast foods. (RW) Line 8: Nowadays, fast foods restaurants **are find (VF) everywhere (SP) (P)** so people eat every (X) food (^). Line 9: For example, if people want to eat pizza, they find many restaurants or they buy **in (PR) (^)** market. Line 10: In addition, some people can be **vegeterian (SP)** so they **can (WW)** every food types (P) but now they find **vegeterian (SP)** foods with these restaurants (RW)

BODY PARAGRAPH 2: Line 11: The second important cause is **provide (VF)** job opportunities. Line 12: Fast food restaurants are very popular because people can have new jobs with these restaurants. (✓) Line 13: For example, nowadays, hamburger is (^) **very (X)** popular

food **types (S/P)**, and many people prefer hamburger **(S/P)**. Line 14: Therefore, people can open specific hamburger restaurants and provide new job (S/P).**(RW)** Line 15: **Other (WF)** example is, Korean cuisine is so popular **last (WW)** years in Turkey because people **wonder (WW) this (PR)** kitchen **(WW)** and they want to try their specific foods. Line 16: Therefore, business people open Korean restaurants and they bring new job opportunities. (✓)

BODY PARAGRAPH 3: Line 17: The third **most (WW)** cause is people can prepare **easily (WO).** **(RW)** Line 18: When people go to work or school, **their (Pr)** don't have enough time **for (PR)** cook something, so fast food restaurants **for (PR)** this **people (S/P)** so good **things (S/P).** **(RW)** Line 19: For example, when people come back home, they order **meal (S/P)** in **(PR)** restaurants. Line 20: Also, they buy in (PR) market everything.**(RW)** Line 21: Fast food restaurants are so popular because people **access (WW)** **(^)** fast and easily.

CONCLUSION: Line 22: People nowadays prefer fast food restaurants in the world because **for (X)** this **have (S/V)** some reasons. Line 23: These causes are people **(^)** access always, have **(^)** new job opportunities, and they can prepare easily. **(RW)** Line 24: If people gain time and access every type of foods, they prefer fast food restaurants **(RW)**(P)**

2.Sample Structured Review Table Sample completed by a teacher for refining error-coded feedback (T6, paper 3)

Note: The teacher sample below shows the structured review table used by a teacher to document changes made to AI-generated error-coded feedback. It includes the raw data of AI-generated error-coded feedback and error descriptions, data used for analyzing the acceptance levels of each piece of AI feedback, written teacher notes for edits, and selected rationale codes.

Section& location	Error Code	Error description by AI	Sentence with error code	AI feedback (Accepted/Not Accepted/Partially Accepted)	Teacher Edits/Changes/Removals	rationale for teacher changes	Additional Rationale (If "Other")
TITLE	(^)	Consider adding a title to your essay. How could you summarize your main topic in a few words?	N/A	accepted (no change needed)		no 15	

INTRODUCTION LINE 1	- (VF)	Think about the verb form of "are increase". How do we usually express continuous actions?	In the world, fast food restaurants are increase (VF) day by day.	accepted (no change needed)		no 15	
INTRODUCTION LINE 2	- (X)	Consider if "are" is necessary here. How do we usually form negative sentences with "don't"?	Fast food restaurants are (X) don't give up, and they always find new food types.	accepted (no change needed)		no 15	
INTRODUCTION LINE 4	- (X, X)	Think about whether "are" and "the" are necessary in this sentence. How do we usually talk about things happening for the first time?	Also, fast food restaurants are (X) started in the U.S.A the (X) first time.	partially accepted (some changes needed)	RW for the underlined part: should be:for the first time	no 8, no 11	
INTRODUCTION LINE 5	- (WO)	Consider the word order in this question. How do we usually form questions in English?	Why (WO) these restaurants most popular?	rejected (major changes needed)	AI didn't give feedback for the missing verb. The student may just change the order of the words without the main verb; which gives the student a wrong message.	no 2 , no 4	
INTRODUCTION LINE 6	- (WW)	Is "tell" the word you want to use here? Think about a word that means "give" or "present".	People prefer these restaurants because they tell (WW) three causes for this topic: These causes are accessing easily, providing job opportunities and preparing easily.	partially accepted (some changes needed)	The student is supposed to write a more proper thesis statement in order to make his/her essay more academic. It needs full rewriting. RW code overlooked here.	no 2 , no 11	

BODY 1 - LINE 7	(WW, WW, X)	Think about whether you need "most" here. Is "people" necessary? Do we need "the" before "fast foods"?	The first most (WW) cause people (WW) is people can easily the (X) fast foods.	rejected (major changes needed)	The correct sentence should be " The first and the most important cause is easy access to fast food." Even if students take errors codes given into considerations, the sentence will still be meaningless.	no 2 , no 11	
BODY 1 - LINE 8	(VF, SP)	Consider the correct form of "are find". How do we usually spell "everywhere"?	Nowadays, fast foods restaurants are find (VF) everywhere (SP) so people eat every food.	partially accepted (some changes needed)	There are some overlooked erros and the second part of the sentence is a little bit vague.	no 8, no 11	
BODY 1 - LINE 9	(PR)	Think about the preposition used with "market". Do we say "buy in market" or something else?	For example, if people want to eat pizza, they find many restaurants or they buy in (PR) market.	accepted (no change needed)		no 15	
BODY 1 - LINE 10	(SP, WW, SP)	Check the spelling of "vegeterian". Is "can" the word you want to use here? Check the spelling of "vegeterian" again.	In addition, some people can be vegeterian (SP) so they can (WW) every food types but now they find vegeterian (SP) foods with these restaurants.	rejected (major changes needed)	The underlined part of the sentence obscures the meaning and should be rewritten. Also, there is a missing verb problem.	no 2, no 4	
BODY 2 - LINE 11	(VF)	Think about the form of "provide" here. What form do we use after "is"?	The second important cause is provide (VF)job opportunities.	accepted (no change needed)		no 15	

BODY 2 - LINE 12		<i>NO ERROR DESCRIPTION</i>		rejected (major changes needed)	Errors provided correctly in word doc, but No error description provided.It lacks guiding tone for correcting error.	no 6	
BODY 2 - LINE 13	(X, S/P)	Consider if "very" is necessary here. Should "types" be singular or plural?	For example, nowadays, hamburger is very (X) popular food types (S/P), and many people prefer hamburger.	accepted (no change needed)		no 15	
BODY 2 - LINE 14	(S/P)	Think about whether "job" should be singular or plural here.	Therefore, people can open specific hamburger restaurants and provide new job (S/P).	partially accepted (some changes needed)	The underlined part should be " ...opening hamburger restaurans will provide people with some job opportunities."	no 8, no 11	
BODY 2 - LINE 15	(WF, WW, WW, PR)	Consider the form of "Other". Is "last" the word you want to use for recent years? Is "wonder" the right word here? Do we say "wonder this kitchen" or something else?	Other (WF) example is, Korean cuisine is so popular last (WW) years in Turkey because people wonder (WW) this (PR)kitchen and they want to try their specific foods.	accepted (no change needed)		no 15	
BODY 3 - LINE 17	(WW, WO)	Think about whether you need "most" here. Consider the placement of "easily" in the sentence.	The third most (WW) cause is people can prepare easily (WO).	rejected (major changes needed)	The student should fix the structure of the sentence and sentence needs rewriting. Also, the students should be reminded that he/she already used -the most important reason-above.	no 2, no 4	

BODY 3 - LINE 18	(Pr, PR, PR, S/P, S/P)	Check if "their" is the right pronoun. Think about the preposition used with "cook". Is "for" the right preposition before "this people"? Should "people" and "things" be singular or plural?	When people go to work or school, their (Pr) don't have enough time for (PR) cook something, so fast food restaurants for (PR) this people (S/P) so good things (S/P).	rejected (major changes needed)	The most important problem about the missing verb was overlooked. There are some problems about the sentence structure and it needs rewriting.	no 2 , no 4	
BODY 3 - LINE 19	(S/P)	Consider whether "meal" should be singular or plural.	For example, when people come back home, they order meal (S/P) in restaurants.	accepted (no change needed)		no 15	
BODY 3 - LINE 20	(PR)	Think about the preposition used with "market". Do we say "buy in market" or something else?	Also, they buy in (PR) market everything.	rejected (major changes needed)	There is a serious problem about coherence in this sentence and it needs full rewriting. The meaning was ignored and RW code was not assigned by AI, only a minor mistake was considered.	no 2, no 4	
BODY 3 - LINE 21	(WW)	Is "access" the word you want to use here? Think about a word that means "get to" or "reach".	Fast food restaurants are so popular because people access (WW) fast and easily.	accepted (no change needed)		no 15	
CONCLUSION - LINE 22	(X, S/V)	Consider if "for" is necessary here. Does "have" agree with "this"?	People nowadays prefer fast food restaurants in the world because for (X) this have (S/V) some reasons.	accepted (no change needed)		no 15	

CONCLUSION LINE 23	- (^, ^)	Think about what might be missing before "access" and "have". Do we need a subject or a verb here?	These causes are people (^) access always, have (^) new job opportunities, and they can prepare easily.	rejected (major changes needed)	The thesis statement should be paraphrased properly from scratch. It needs full rewriting, but the RW code was not assigned by AI for this situation.	no 2, no 4	
CONCLUSION LINE 24	- (P)	Consider what punctuation might be needed at the end of this sentence.	If people gain time and access every type of foods, they prefer fast food restaurants**(P)**	rejected (major changes needed)	The meaning of this sentence is obscured. The student should rewrite it to make this sentence meaningful.	no 2 , no 4	

3. Structured Review (Content Feedback)

Note: The teacher sample below shows the structured review table used by a teacher to document their suggestions for the content-feedback changes made to AI-generated feedback. It includes the raw data of AI-generated content feedback, data used for analyzing the acceptance level, written teacher notes for suggested changes, and selected rationale codes.

<i>Areas needing improvement</i>	<i>AI feedback (Accepted/ Not Accepted/Partially Accepted)</i>	<i>Teacher Edits/Changes/Removals- Write where you will edit (strenghts, areas needing improvement, or both, and write your change(s) if there is any.</i>	<i>Choose the number from drop-down list for your rationale for Teacher Changes</i>	<i>Additional Rationale (If "Other")</i>
Areas for Improvement: 1. Thesis Statement: Try to make your main points clearer in the introduction. For example, you could write: "People prefer fast food restaurants for three main reasons: they are easily accessible, they provide job opportunities, and they offer convenient meal options." 2. Supporting Details: In each body paragraph, consider adding more specific examples: o In the accessibility paragraph, can you mention	partially accepted (some changes needed)	The first and third supporting ideas for the body paragraphs are similar. It would be better if the student use a different reason for the third body paragraph. For ex: being affordable The feedback given for the thesis statement is too direct.	no 19, no 22	

specific ways fast food is easily available (e.g., delivery apps, drive-throughs)?

- o In the job opportunities paragraph, try to provide some statistics about employment in the fast food industry.
- o In the convenience paragraph, explain more clearly how fast food fits into people's busy lifestyles.

3. Language Use: Pay attention to verb forms and subject-verb agreement. For example, "Fast food restaurants are increase" should be "are increasing." Also, try to use more varied vocabulary. Instead of repeating "fast food" so often, you could use phrases like "quick-service restaurants" or "convenience meals."

APPENDIX-11. Qualitative Analysis of Teachers' error-coded feedback, and Representative Quotes for RQ1

1. Qualitative Analysis of Teachers' error-coded feedback reviews

Note: Teachers' review notes helped support RQ1 findings with qualitative data to contextualize their refinements of AI-generated error-coded feedback.

1. Overlooked Rewriting (RW) Code

Paper #	Student Text	Teacher Comments/Representative Quotes	Issue
2	Since increasing fast food restaurants, we tend to be lazy, but we should stop it and doing (VF)...	"RW code should be added to the first clause; 'doing' should be 'prepare'." (Paper 2/ T4) "added RW code as the beginning of the sentence with 'since' wasn't completed and added WW code for 'doing' it is also a wrong word to use here." (Paper 2/ T2)	Sentence Fragment
2	Although (TR) we have knowledge about includes (WF) in traditional foods.	"This is not a full sentence and should be rewritten, but RW code was not provided." (Paper 2/T1) "Incomplete sentence fragment. There is a need for rewriting." (Paper 2/T5)	Sentence Fragment
5	It is such a problem that (WW) mentally or physically.	"This is not a full sentence, but there is no feedback for it. AI cannot understand what feedback to give when there is no full sentence or [when a] sentence [is] incomplete. WW code given was not fixing anything." (Paper 5/T5)	Sentence Fragment
4	In (SP) my opinion 2 (WF) causes of growing role of using robots.	"Incomplete sentence fragment (there are) missing, there is a need for rewriting." (Paper 4/T5)	Sentence Fragment
3	People prefer these restaurants because they tell (WW) three causes for this topic: These causes are accessing easily, providing job opportunities and preparing easily.	"Ss needs to rewrite a better [thesis] statement, so he [/]she needs to change the whole sentence." (Paper 3/T1) "Although the student had many mistakes, RW code wasn't given. It's necessary for [parallelism]." (Paper 3/T4)	Poor thesis structure
7	Foods are it canbe (WW) wasting and loss effects the (A) decrease (SP) popularity of traditional food restaurants.	"The sentence has a serious problem about structure and it is grammatically incorrect and it is not a proper thesis statement either. RW code was not provided to these cases oftentimes. This can be a nuance it falls short of." (Paper 7/T6)	Poor thesis structure

8	So there are three effects of increasing popularity of fast food restaurants; be (VF) obeze (SP), spend money and unhealthy (WF).	"Thesis needs to have parallel form in verbs, but AI could catch this nuance. I suggest VF code for 'spend'. 'Spend' should be 'spending'. Wf code for unhealthy is unnecessary code." (Paper 8/T6)	Poor thesis structure
8	The second effect is spent (VF) money because people need to (X) a lot of money	"The topic sentence is vague. needs rewriting." (Paper 8/T6)	Poor Topic sentence structure
5	I know I mentioned 1 positive side and two negative sides, but it has (WW) not to be all negative or positive every time.	"Tone is not academic and meaning unclear. RW needed." (Paper 5/T6)	Tone Problem
10	This is a problem as I wrote last (WW) paragraphs and we must solve this problem.	"This is not academic way of summarizing ideas." - (Paper 10/T4)	Tone Problem
5	Fast food restaurants be (VF) a barrier (SP) against to (X) them and do not help their self improves (WW).	"not a meaningful sentence. RW code is necessary here. AI could not see the sentence is not clear in meaning: barriered against what?/ health? How is this helping personal improvement?/ not meaningful at all." (Paper 5/T5)	Clarity of Meaning
8	In (SP) summary, life is short but your body is very strong and long.	"but your body is very strong and long (RW for clarity). The grammar can be correct, but what about meaning?" (Paper 8/T2)	Clarity of Meaning
10	If we continue to use robots more than people in our companys (SP), This (C) can create a grup (SP), then they establish companys (SP) and this situation geting (SP) bigger and finally they can start a war.	"AI tries to clear up basic problems, but it missed the pronoun reference errors and therefore the sentence does not make sense. It is not clear what 'they' refer to. And also it is not clear what 'this situation' refers. I would ask student to make them clear: 'Then they establish companies': Who is 'they'? The robots? 'This situation getting bigger': What exactly is growing?" - (Paper 10/T4)	Clarity of Meaning -pronoun reference error
8	Especially, McDonald's (P) Domino's (P) Burger King...(P) people need food but it isn't fast food.	"it isn't fastfood needs rewriting. AI should help student to reconsider what he/she means for clear expression" (Paper 8/T2)	Clarity of Meaning
9	In (SP) addition, people are (X) improved themselves (SP) for their jobs by the (X) their robots and they can earn many new information.	"This sentence contains multiple errors, making it difficult to understand. Earn info: gain info... (word choice) improved: improve (tense), but this is still not helping alone. It is best to suggest student to write the sentence again in various errors at once." (Paper 9/T2)	Multiple errors

2. Overlooked Codes

Paper #	Student Text	Teacher Comment	Issue
2	We should be careful about how much eating (VF) fast food...	"Missing subject is a major overlooked problem in my opinion (how much we eat)" – T4 "Major overlooked code for missing subject: 'how much WE eat'" – (Paper 2/T2)	Overlooked code for missing Subject
2	Maybe fast food is a (A) opportunity for using more time but most people consume fast food because of (X) it is easy.	"Two major overlooked mistakes regarding word choice overlooked and this can affect understanding of sentence if not corrected. It should be 'using time effectively' and 'it is easy to order'" (Paper 2/T6)	Overlooked code for word choice
5	People's Choice (C) affect them and their lifestyle shapes according to their decision.	"Choice is not just C problem, but also S/P: choices. For the passive form, no VF feedback. The correct form should be 'are shaped'" (Paper 5/T7)	Overlooking codes when more than one error exists for the same word

3. Incorrect Coding

Paper #	Student Text	Teacher Comment	Issue
10	In conclusion, robot's (WF) number is increasing and this have got (VF) several effects for our lives.	"Wf and VF code are both incorrect codes. Have got should be has got, so S/v would help to match the verb with subject: this." (Paper 3/T10)	Wrong code
4	Because these companys (SP) are makes (VF) new items and learing (SP) to new ideas with people.	"instead of SP, WW code should be given. major mistake for AI" (Paper 4/T2)	Wrong code
2	Our metabolism being (VF) slower so we can gaining (VF) weight easily	"'being' must be changed to 'get' slower. Rather than a verb form, it is wrong word choice, and AI did not understand it. There must be a comma before so, but no feedback provided." (Paper 2/T1)	Overlooked punctuation, misidentified error type
5	"They always eat and eat"	"AI marked 'They always eat and eat' as correct, but the phrase is awkward and unnatural. It likely passed because AI focused on grammatical correctness rather than natural usage. The RW code was added for better phrasing" (Paper 5/T8)	Incorrect sentence, but ticked as correct
6	By trying to make producing and working faster and better, scientists put robots to the job and remove humans from their job (✓)	"AI ticked the sentence, but 'producing' (WF) should be production. And also, 'put the job' is not natural, job is WW, and should be 'put to work.'" (Paper 6/T7)	Incorrect sentence, but ticked as correct

4. Redundant Coding

Paper #	Student Text	Teacher Comment	Issue
8	People should be healthy (WF) body, doing (VF) sport and don't (S/V) eat unhealthy food.	"I added WW code for the word 'be' which should be verb: have. The code for healthy is redundant." (Paper 8/T7)	Redundant Code
7	Second (WF) important effect the (WW) waste of food.	"unnecessary code was given for 'second' and also there is a missing main verb (is). AI should have highlighted that part by giving missing word code." - (Paper 2/T2)	Redundant code, plus missing code for missing verb
2	Cooking traditional foods more (WF) difficult than order (VF) a (X) food from fast food restaurants.	"redundant code: WF for 'more'. 'More' is correct. Cooking traditional foods (^), here 'verb' is missing and this is major overlooked code for me." (Paper 2/T5)	Redundant code for "more"

APPENDIX-12: Content Feedback Analysis Table and a sample of coding

1. Content Feedback Analysis Table results

Content Feedback Analysis Table				
Category	Subtheme	Item	Count	Percentage
Content Development Feedback (CDF)	Unaddressed Ideas	Repetitive		
		Repetition of ideas Not Addressed for body paragraphs	34	27.6%
		Repetition of ideas Not Addressed for topic sentences	1	0.8%
		No feedback for repetitive ideas in thesis statement	1	0.8%
		Subtotal	36	29.3%
	Missing Structural Feedback			
		Missing feedback for the conclusion	7	5.7%
		No feedback for the hook sentence missing in the introduction	1	0.8%
		Lack of Thesis Statement not mentioned	2	1.6%
		Lack of feedback on thesis/topic sentences for emphasizing cause/effect	1	0.8%
		Subtotal	11	8.9%

	Missed or Insufficient Feedback for Body Paragraph Development			
		Expansion problem is not addressed	5	4.1%
		Expansion problem caused by length of paragraph is not addressed	1	0.8%
		Circular reasoning not addressed within paragraphs	2	1.6%
		Needs More Actionable Suggestions	3	2.4%
		Unnecessarily given/redundant feedback for the supporting idea	1	0.8%
		Misinterpretation of student idea and missed feedback for developing student's idea	2	1.6%
		Robotic/mechanic feedback lacks in depth for improving the student's paragraphs	1	0.8%
		Subtotal	14	11.4%
		CDF Total	61	49.6%
Direct Feedback (DF)				
		Direct Thesis Feedback without guidance	25	20.3%
		DF Total	25	20.3%
Feedback Tone (FT)				
		Lacks reflective/guiding feedback to help avoid repetitive ideas	9	7.3%
		Lacks strength and clarity in guiding the student to add supporting ideas	2	1.6%

		Lacks reflective/guiding feedback that encourage student to clarify his/her focus	2	1.6%
		FT Total	13	10.6%
Missing Feedback Error (MFE)				
		Error marked in Content feedback, missed in error code feedback	4	3.3%
		MFE Total	4	3.3%
Missing Feedback – Language Use (MF-LU)				
		Lack of suggestions for Language Use points	1	0.8%
		Lack of feedback on the cause & effect language use	1	0.8%
		MF-LU Total	2	1.6%
Positive Feedback (PF)				
		Good suggestions for Language Use points	6	4.9%
		Good suggestions for supporting ideas	12	9.8%
		PF Total	18	14.6%
		Grand Total	123	100.0

2.Sample of the Content-feedback's Analysis Conducted by the researcher (a sample coding scheme for T3)

Note: This includes highlighted sections, and preliminary and descriptive codes inducted from the data.

	T3				
PAPERS	RATIONALE CODES	Teacher notes for content feedback changes	Preliminary Codes	Descriptive codes	My Comments
PAPER 2	no 23				
PAPER 3	no 19	The first and the third body paragraphs look pretty much alike. We NEED a different reason.	*Repetition of ideas Not Addressed for body paragraphs	CDF-Content Development Feedback Issues	
PAPER 4	no 23	I could use it after leaving half of the thesis statement blank for students to complete.	*Direct Thesis Feedback without guidance		
PAPER 5	no 19, no 22	There are A LOT of repetitions. The essay is not expanded well. Circular reasoning: "people are obese, they eat fast food, they're obese 'cause they eat fast food. Fast food makes you obese, so you are obese, because you eat fast food." Paragraph 1&2 suffer from this circular reasoning, meanwhile prg.3 is full of mistakes and repetitions.	*Repetition of ideas Not Addressed for body paragraphs * Circular reasoning not adressed within paragraphs	CDF-Content Development Feedback Issues	Fails to identify or respond to circular logic in BD1 & BD2, where reasoning repeats without development
PAPER 6	no 23	I accept this content FB (feedback) but similar to paper 5, I do not see any FB on repetitions.!!!	*Repetition of ideas Not Addressed for body paragraphs	CDF-Content Development Feedback Issues	
PAPER 7	no 19, no 22	There are A LOT of repetitions. The essay is not expanded well. It is similar to "paper 5."	*Repetition of ideas Not Addressed for body paragraphs * Expansion problem is not adressed	CDF-Content Development Feedback Issues	
PAPER 8	no 19, no 22	It's ok but I still do not see any FB about repetitions!!! Obesity and health issues are repeated in the body parts.	*Repetition of ideas Not Addressed for body paragraphs	CDF-Content Development Feedback Issues	(e.g. Teacher noted, obesity and health paragraphs)
PAPER 9	no 19, no 22	1. I don't think there's a good hook sentence available, and there's no FB on this. 2. There seems repetitive words in the thesis statement, e.g., "robots", and "people".	*No feedback for the hook sentence missing in the introduction *No feedback for repititive ideas in thesis statement	CDF-Content Development Feedback Issues	
PAPER 10	no 19, no 22	Thesis is also direct and it does not allow for self-correction.AI commented: Job loss and job displacements paragraph. Actually, the ideas are very similar, especially topic sentences. Maybe changing wordings would help, but AI did not provide any feedback on this. The rest of the ideas could be useful.	*Direct Thesis Feedback without guidance *Repetition of ideas Not Addressed for topic sentences	*DF:Overly direct feedback issues *CDF-Content Development Feedback Issues	(Teacher noted,maybe AI asking to change wording of the topic sentences of bd1& 2 to help student change their focus and differentiate the paragraphs)

APPENDIX-13: Qualitative Analysis of Teachers' content-feedback reviews, and Representative Quotes for RQ1 (highlighted)

Note: Teachers' review notes served as qualitative support to help contextualize content refinements of AI-generated error-coded feedback. Only the highlighted quotes were used in the thesis analysis, while the remaining ones were included to ensure comprehensive data presentation with frequency information.

Category	Subcategory (Count, %)	Teacher Quotes
1. Unaddressed Repetitive Ideas (36, 29.0%)	Repetition in body paragraphs (34, 27.4%)	"Obesity and the health issues are basically the same things. The student didn't get feedback about it, which has a serious effect on content." (Paper 8/T6) "The ideas in the first and second body paragraphs about robots replacing human beings are repetitive, yet no feedback was given for this repetition." (Paper 6/T2) "I think the subjects for the first and second bodies are nearly identical in student writing. They can be merged in a single paragraph with concrete examples."(Paper 5/T4) "This student needs better guidance for body paragraphs 1 and 3. The ideas are quite identical, though the student is trying to present two consequences: one is replacing their jobs, and the other is losing their jobs." (Paper 10/T4) "The first and the third reasons are similar. It would be better if the student find a complete new reason for the 3rd paragraph such being affordable." (Paper 3/T1) "The most problematic thing about the essay there are no completely different 3 ideas for body paragraphs." (Paper 4/T1) "Body 1 and body 2 ideas and examples are similar. This was not mentioned in content feedback part." (Paper 6/T1) "The first two ideas can be combined as they are so close to each other and student can find a 3rd new idea." (Paper 5/T1) "The feedback on repetitive supporting ideas in body 1 and 3 was not mentioned in AI feedback." (Paper 3/T5) "The content of body 1 and 3 are basically same, but the student did not get feedback about it." (Paper 10/T6) "Obesity and health issues are basically the same thing. There is no feedback about it and this has a serious effect on content." (Paper 8/T1) "The feedback on repetitive supporting ideas in body 1 and 3 was not mentioned in AI feedback." (Paper 3/T2) "The ideas in the first and second paragraphs are quite similar. There is a lot of repetition in the essay." (Paper 5/T2) "The first and the third body paragraphs look pretty much alike. We NEED a different reason." (Paper 3/T3) "There are A LOT of repetitions. The essay is not expanded well." (Paper 5/T3) "I do not see any FB on repetitions!" (Paper 6/T3) "There are A LOT of repetitions." (Paper 7/T3) "I still do not see any FB about repetitions! Obesity and health issues are repeated in the body parts." (Paper 8/T3) "Although

		body 1 and body 3 paragraphs are similar to each other in terms of content and ideas. I couldn't see any comments related to this issue from AI's feedback." (Paper 3/T4) "The first and second paragraphs are identical in terms of repeating the idea 'replacing human beings'." (Paper 6/T4) "I think feedback for body 2 and 3 is ok, but for body 1, if student wrote about health issues, then why, student wrote body 2. It would be too repetitive to mention health issues in both bodies." (Paper 5/T5) "The first and second body paragraph ideas about robot replacing human beings are repetitive, and no feedback provided for this repetition." (Paper 6/T5) "AI should also ask the student to compare body paragraphs 1 and 3 for similarity in content." (Paper 10/T5) "The first and third supporting ideas for the body paragraphs are similar." (Paper 3/T6) "Students ideas for the main body paragraphs are so close to each other." (Paper 4/T6) "The topics used for the 1st and 2nd body are so close, so they can be combined in one paragraph with relevant examples." (Paper 5/T6) "The student should have got feedback about the content of the 1st and 2nd body since they are pretty close to each other." (Paper 6/T6) "There is no feedback about the most serious problem which is the lack of three different ideas." (Paper 10/T6) "The ideas are very similar for the main bodies." (Paper 4/T7) "I would prioritize Body Paragraph 1 and Body Paragraph 2, as they present similar ideas." (Paper 5/T7) "It should also prompt the student to examine body paragraphs 1 and 3 for content similarity." (Paper 10/T7) "The first and the third body paragraphs look alike." (Paper 3/T8) "For body paragraphs 1 and 2, I generally understood the student's ideas, but they should expand on them using different supporting details." (Paper 5/T8) "The body paragraphs 1 and 3 lack differentiation." (Paper 10/T8)
	Repetition in topic sentences (1, 0.8%)	"AI commented: Job loss and job displacements paragraph. Actually, the ideas are very similar, especially topic sentences. Maybe changing wordings would help, but AI did not provide any feedback on this." (Paper 10/T3)
	Repetition in thesis statements (1, 0.8%)	"There seems repetitive words in the thesis statement, e.g., 'robots', and 'people'."(Paper 9/T3)
2. Missed or Insufficient Feedback for Body Paragraph Development (14, 11.4%)	Expansion problem not addressed (5, 4.1%)	"The other important problem is ignoring the expansion in the content for body 1 and 2, which makes this essay unbalanced." (Paper 7/T6) "The student should also receive comments on the length of body paragraph 1, as it requires expansion due to a lack of supporting details." (Paper 10/T1) "There is no feedback about lack of support for body 1 and body 2 which is a serious problem." (Paper 7/T1) "The essay is not expanded well." (Paper 7/T3) "The student should get feedback about the length of body 1. The content should be expanded since there is almost no support." (Paper 10/T6)

	Needs More Actionable Suggestions (3, 2.4%)	"The suggestions for the body paragraphs are quite useful, but maybe could be even more effective with slightly more specific actionable ideas... Student may need a suggestion of an idea like that, otherwise, they may read it, but not apply it." (Paper 7/T7) "As for body 1, the idea suggested can be useful, but I think I would ask student to expand on his/her ideas because the phrases are like unfinished and left unsupported such as 'it isn't fast food'. So what, why not fast food? Wanting from student completely new thing may be hard for student to incorporate into his/her writing, especially if student does know about the obesity rates." (Paper 8/T4) "For body 3 though, the suggestion could have been different because student already wrote examples of emotional support and entertainment. If AI wants more, it should have been more clear then." (Paper 9/T7)
	Circular reasoning not addressed (2, 1.6%)	"Circular reasoning: 'people are obese, they eat fast food, they're obese 'cause they eat fast food.' Paragraph 1&2 suffer from this circular reasoning." (Paper 5/T3) "The reasoning is circular: 'people are obese, they eat fast food, they're obese because they eat fast food. Fast food makes you obese, so you are obese because you eat fast food.' Paragraphs 1 and 2 exhibit this circular reasoning." (Paper 5/T2)
	Misinterpretation of student idea (2, 1.6%)	"The content feedback given for the 3rd body paragraph is misleading for the student because student meant how fast food consumption has created a sector and she/he is not talking about marketing strategies." (Paper 5/T6) "There is a serious content feedback problem for the 3rd body paragraph. Student meant how fast food consumption created a sector and student is not talking about marketing strategies but there is no feedback about it." (Paper 5/T1)
	Unnecessarily given feedback (1, 0.8%)	"For the companion paragraph, I think student somehow explained how. So, maybe because AI felt it had to provide a comment, it provided a comment." (Paper 9/T8) "I think AI should be trained to think what the student paper needs, and if ideas for body is enough, there is no reason of wanting more." (Paper 9/T8)
	Robotic/mechanic feedback (1, 0.8%)	"The content feedback is not enough to make students improve the body paragraphs. The content feedback is so robotic." (Paper 8/T6)
3. Missing Structural Feedback (11, 8.9%)	Missing conclusion feedback (7, 5.7%)	"Feedback for conclusion should be given to the students because restatement of ideas is missing which is a major mistake. Final thoughts should be added." (Paper 4/T2) "AI did not check whether the conclusion effectively restates the main points of the essay." (Paper 10/T7) "Actually, this issue appears in all the essays so far." (Paper 10/T5) "Lastly, there is no comment about conclusion. There should

		be a comment about restatement of the ideas." (Paper 4/T4) "Conclusion feedback is needed. (restatement of ideas is missing, and add some final thoughts)." (Paper 4/T5) "I think here conclusion for feedback is missing. The rest of feedback was useful, but in other papers as well I do not see feedback for conclusion." (Paper 3/T7) "Thesis is direct and not allowing for self-correction. No feedback for conclusion." (Paper 4/T8)
	Missing thesis statement feedback (2, 1.6%)	"The student doesn't have a problem with the accuracy of thesis statement, there is no thesis statement at all." (Paper 7/T6) "Instead of mentioning it and guiding the student to write, thesis statement is given." (Paper 7/T5) "There is no feedback about the lack of thesis statement." (Paper 7/T1)
	Missing cause/effect emphasis (1, 0.8%)	"I think the biggest problem is in the thesis. Yes, it is grammatically correct, but AI should provide feedback to help them categorize their ideas into negative effects and positive effects." (Paper 5/T7) "The same applies to the topic sentence." (Paper 5/T7)
	Missing hook feedback (1, 0.8%)	"I don't think there's a good hook sentence available, and there's no FB on this." (Paper 9/T3)
4. Direct Feedback (25, 20.3%)	Providing direct thesis rather than guiding (25, 20.3%)	"The thesis statement was given directly, which can prevent productivity and originality of the students." (Paper 6/T4) "Direct thesis statement was provided by AI and this prevents student from self-correcting their mistake." (Paper 1/T7) "I could use it after leaving half of the thesis statement blank for students to complete." (Paper 4/T3) "There is direct feedback. It directly gives the correct sentence." (Paper 2/T1) "The feedback for the thesis is directly instead of encouraging the student to find it out." (Paper 3/T1) "There is direct feedback for thesis statement." (Paper 6/T1) "There is a direct thesis provided." (Paper 10/T1) "The only issue in the overall feedback is that the thesis statement was rewritten for the student, so I accepted this." (Paper 2/T2) "The issue in the overall feedback is that the thesis statement was rewritten for the student." (Paper 3/T2) "The student was provided with a direct thesis statement solution." (Paper 6/T2) "Thesis is also direct and it does not allow for self-correction." (Paper 10/T3) "I don't think that it is a good idea to give a direct thesis statement sentence. The student can copy it directly." (Paper 4/T4) "Furthermore, the thesis statement was given directly, which can prevent productivity and originality of the students." (Paper 6/T4) "I liked all the points except for the thesis statement. It should not be directly given." (Paper 7/T4) "The thesis suggestion is fine, but instead of giving the exact way to write it, AI should guide the student to rewrite it with a clear statement of three ideas." (Paper 10/T4) "The correction for the thesis should not be directly provided, but the rest of feedback is beneficial for the student." (Paper 2/T5) "Thesis statement was given direct correction." (Paper 3/T5) "There was direct

		thesis statement solution for the student." (Paper 6/T5) "There are some direct feedback for the thesis statement and language use." (Paper 2/T6) "The feedback given for the thesis statement is too direct." (Paper 3/T6) "Firstly, the example for the thesis is too direct than it is supposed to be." (Paper 5/T6) "Also, the suggestion for the thesis statement is too direct than it is supposed to be." (Paper 6/T6) "Also, the suggestion given for that problem is too direct." (Paper 7/T6) "The feedback given for the thesis statement is so direct." (Paper 10/T6) "Direct thesis statement was provided by AI and this prevents student from self-correcting their mistake." (Paper 2/T7) "Thesis is direct and not allowing for self-correction." (Paper 4/T8)
5. Feedback Tone (13, 10.6%)	Lacks reflective/guiding feedback for repetitive ideas (9, 7.3%)	"The student should get content feedback about choosing different ideas for different paragraphs to improve his/her essay." (Paper 5/T6) "It would be too repetitive to mention health issues in both bodies... student was not told that both bodies are similar and student should not use same points." (Paper 5/T5) "To differentiate them, the student needs to explain in Body Paragraph 1 how people become inactive, less energetic, or feel fatigued." (Paper 5/T7) "The student should think about whether these are truly different ideas or just repeating the same point in different words. If they are distinct, how can they be clearly separated in the essay? If they overlap too much, would it be better to combine them into one paragraph?" (Paper 10/T4) "The body paragraphs 1 and 3 lack differentiation, and AI fails to provide reflective feedback to help the student recognize this. It should invite student to think about this and either combine two paragraphs or enrich these bodies with more ideas to make student points more clear." (Paper 10/T2) "AI should also ask the student to compare body paragraphs 1 and 3 for similarity in content and suggest that the student use distinct supporting details if the paragraphs are not exactly the same. If necessary, AI should ask the student to combine body paragraphs that present the exact same idea." (Paper 10/T5) "If the paragraphs convey the same point, AI could suggest combining them. Providing both suggestions would allow the student to choose the best approach." (Paper 10/T7) "This is a little bit confusing (vague for student in a way that student may not know how to use this feedback) because student was not told that both bodies are similar and student should not use same points." (Paper 5/T5) "AI provided some supporting ideas related to an unhealthy lifestyle, but if the student mentions diseases, it may not effectively differentiate body 1 and body 2 from each other." (Paper 5/T8)
	Lacks strength/clarity for supporting ideas (2, 1.6%)	"But it could also explain why specific examples are needed, since the student might not think they are really necessary or must be included. I am not sure if student take action after reading the

		feedback." (Paper 2/T4) "The AI suggested adding more specific examples, but it should have given a stronger suggestion... Otherwise, these paragraphs are not completed from my perspective." (Paper 8/T2)
	Lacks guidance to clarify focus (2, 1.6%)	"As for the marketing paragraph, I would ask the student to clarify its focus." (Paper 5/T7) "I assume this paragraph discusses the positive effects on marketing, but this is not explicitly clear throughout the paragraph." (Paper 5/T7)
6. Missing Feedback Error (4, 3.3%)	Errors marked in content feedback but missing in error code (4, 3.3%)	"In the language use part of content feedback, there is feedback about the verb but there is no error code in the related part in the word file." (Paper 4/T1) "The mistake in line 2 was corrected in content feedback, but it was not corrected in error code feedback; this shows that AI cannot give the respective code." (Paper 5/T5) "In line 11, AI marked the sentence as correct, but there is a missing verb problem and this mistake was not included in the error code list." (Paper 4/T6) "The mistake in line 2 was corrected in content feedback however it wasn't corrected in error code feedback this shows that AI feedback cannot give two error codes respectively for the same word." (Paper 4/T2)
7. Language Use Feedback (2, 1.6%)	Lack of suggestions for Language Use points (1, 0.8%)	"I think AI can suggest that the student consider his/her word choices to make their ideas clearer." (Paper 5/T8)
	Lack of feedback on cause & effect language (1, 0.8%)	"I would advise the student to use cause-effect language, as I do not see words like 'lead to' or 'cause', which are essential for clarity." (Paper 5/T7)
8. Positive Feedback (18, 14.6%)	Good suggestions for supporting ideas (12, 9.8%)	"I would definitely say that students needs to write in more clear way and support his/her body paragraph with more examples because the ideas are written without any specific examples. I think AI also asks for specific examples and that is good." (Paper 6/T7) "The ideas for the body paragraphs are quite useful and very similar to what I would give because AI identified the gaps and provide targeted feedback for the supporting ideas, especially comment for body 2 and 3 is straight to the point." (Paper 8/T4) "The suggestions for body paragraphs are actually useful, especially for body 3, it makes clear that student need to write more to support his/her idea." (Paper 7/T8) "The suggestions for bodies are all relevant, and to the point, for instance, the last paragraph of the student says, 'we need to be careful', but as AI noted, in the first place, the situation could have been explained better." (Paper 6/T8) "I would correct nothing. I found all the suggestions for body useful and they all cater for student needs." (Paper 8/T7) "I found all the suggestions for body useful and they all cater for student needs." (Paper 9/T7) "I really liked the

		comment for body 1 because I was thinking student is not clearly saying what makes robots practical. Same for body 2. Student said robots help people in various jobs, but what jobs? Student should specify them." (Paper 9/T8) "I liked all the points except for the thesis statement. It should not be directly given. Comments for body paragraphs are detailed and easy to take action for the student." (Paper 7/T4) "Details suggested for increasing examples for the paragraph are useful, too." (Paper 2/T4) "AI has suggested good points in terms of supporting usage to improve the quality of the essay." (Paper 4/T4) "The suggestions for the body paragraphs are useful for expansion. I definitely agree with the suggestion regarding transition words." (Paper 5/T7) "I would encourage the student to support their points with examples to strengthen their argument." (Paper 5/T7)
	Good suggestions for Language Use points (6, 4.9%)	"The comments for language section for the grammar points and suggestion for varied vocabulary are useful." (Paper 2/T4) "The suggestions for word form and verb form is very useful, as this way, student can understand repeated problems in the essay." (Paper 7/T7) "The repeated mistakes of word form of health was also mentioned in the AI feedback. Also, offering students different vocabulary to use instead of fast food is good idea." (Paper 6/T8) "Content feedback about language use and expansion problems are correct and suitable here." (Paper 7/T2) "AI has suggested good points in terms of language usage to improve the quality of the essay." (Paper 4/T4) "I like the comments provided for language use. I think student can benefit from these, especially, suggesting correct word choice for engineering has growing up." (Paper 4/T8)

APPENDIX-14: Teachers' Accuracy Evaluations

Note: This Appendix covers acceptance level count sample (1), rubric evaluation (full data) (2), reliability check outouts (3-4), and qualitative accuracy evaluations for error-coded feedback (5), and for content feedback (6).

1.A sample Acceptance Levels ratings document (T7, across 9 papers)

AI feedback (Accepted/ Not Accepted/Partially Accepted)	File
accepted (no change needed)	T7 Paper No 2_.xlsx
accepted (no change needed)	T7 Paper No 2_.xlsx
accepted (no change needed)	T7 Paper No 2_.xlsx
accepted (no change needed)	T7 Paper No 2_.xlsx
partially accepted (some changes needed)	T7 Paper No 2_.xlsx
accepted (no change needed)	T7 Paper No 2_.xlsx
partially accepted (some changes needed)	T7 Paper No 2_.xlsx
partially accepted (some changes needed)	T7 Paper No 2_.xlsx
partially accepted (some changes needed)	T7 Paper No 2_.xlsx
partially accepted (some changes needed)	T7 Paper No 2_.xlsx
accepted (no change needed)	T7 Paper No 2_.xlsx
accepted (no change needed)	T7 Paper No 2_.xlsx
accepted (no change needed)	T7 Paper No 2_.xlsx
partially accepted (some changes needed)	T7 Paper No 2_.xlsx
rejected (major changes needed)	T7 Paper No 2_.xlsx
rejected (major changes needed)	T7 Paper No 2_.xlsx
accepted (no change needed)	T7 Paper No 2_.xlsx
partially accepted (some changes needed)	T7 Paper No 2_.xlsx
rejected (major changes needed)	T7 Paper No 2_.xlsx
rejected (major changes needed)	T7 Paper No 2_.xlsx
rejected (major changes needed)	T7 Paper No 2_.xlsx
rejected (major changes needed)	T7 Paper No 2_.xlsx
rejected (major changes needed)	T7 Paper No 2_.xlsx
partially accepted (some changes needed)	T7 Paper No 2_.xlsx
accepted (no change needed)	T7 Paper No 3_.xlsx
accepted (no change needed)	T7 Paper No 3_.xlsx
accepted (no change needed)	T7 Paper No 3_.xlsx
rejected (major changes needed)	T7 Paper No 3_.xlsx
rejected (major changes needed)	T7 Paper No 3_.xlsx
partially accepted (some changes needed)	T7 Paper No 3_.xlsx
rejected (major changes needed)	T7 Paper No 3_.xlsx
partially accepted (some changes needed)	T7 Paper No 3_.xlsx
partially accepted (some changes needed)	T7 Paper No 3_.xlsx
rejected (major changes needed)	T7 Paper No 3_.xlsx

[illegible]

[illegible]

[illegible]

rejected (major changes needed)	T7Paper No 10_.xlsx
rejected (major changes needed)	T7Paper No 10_.xlsx
accepted (no change needed)	T7Paper No 10_.xlsx
rejected (major changes needed)	T7Paper No 10_.xlsx
rejected (major changes needed)	T7Paper No 10_.xlsx
rejected (major changes needed)	T7Paper No 10_.xlsx
accepted (no change needed)	T7Paper No 10_.xlsx
rejected (major changes needed)	T7Paper No 10_.xlsx
rejected (major changes needed)	T7Paper No 10_.xlsx
rejected (major changes needed)	T7Paper No 10_.xlsx

2. Teacher Rubric Evaluations, and commentaries, and Representative Quotes (highlighted)

Paper	Teacher	Rating	Comment
Paper 1	T1	2	"There were correct error codes. However, there were still some sentences which lacked major overlooked errors, and those errors obscure the meaning."
Paper 1	T2	2	"A lot of overlooked codes in this paper. In addition, the code RW was not used in necessary lines which affect the accuracy of the essay and direct the students incorrectly."
Paper 1	T3	2	"Rarely major mistakes, partially overlooked and wrong placements."
Paper 1	T4	2	"I partially accepted the feedback because I felt that there are a lot of overlooked errors not provided any feedback."
Paper 1	T5	2	"There are many overlooked mistakes, with some major mistakes such as wrong error codes."
Paper 1	T6	2	"Although AI gave correct error codes, there are still some major problematic areas overlooked."
Paper 1	T7	2	"Although AI did not provide a lot of wrong codes, it overlooked many errors."
Paper 1	T8	2	"Due to presence of overlooked items."
Paper 2	T1	2	"There is still overlooked and vague feedback. Some of them are too direct."
Paper 2	T2	2	"A lot of overlooked codes in this paper. In addition, the code RW was not used in necessary lines which affect the accuracy of the essay and direct the students incorrectly."
Paper 2	T3	2	"Rarely major mistakes, partially overlooked and wrong placements."

Paper 2	T4	2	"There are so many overlooked items, global errors are missed and feedback is not given."
Paper 2	T5	2	"The same issue: there are numerous overlooked items."
Paper 2	T6	2	"Some suggestions are so direct, so such feedback will not help students think and fix their own mistakes." ★
Paper 2	T7	2	"Same with paper 1, the presence of overlooked codes on the paper led me to think so."
Paper 2	T8	2	"Due to presence of overlooked items."
Paper 3	T1	2	"As a result, it is impossible to use the feedback given without making any changes."
Paper 3	T2	2	"A lot of overlooked codes in this paper. In addition, the code RW was not used in necessary lines which affect the accuracy of the essay and direct the students incorrectly."
Paper 3	T3	2	"Rarely major mistakes, partially overlooked and wrong placements."
Paper 3	T4	2	"There are so many overlooked items, global errors are missed and feedback is not given."
Paper 3	T5	2	"The same issue: there are numerous overlooked items."
Paper 3	T6	2	"A student cannot fix the problems only with the given feedback in some cases." ★
Paper 3	T7	2	"There were so many overlooked codes on the paper and this inhibits the student correction."
Paper 3	T8	2	"Due to presence of overlooked items."
Paper 4	T1	2	"In some parts, there is no feedback on structure while there are problems with meaning."
Paper 4	T2	1	"So many rejected items especially, RW code related problems led me to this decision."
Paper 4	T3	1	"So many overlooked mistakes and errors..."
Paper 4	T4	1	"Rewrite code was not provided although required in many sentences with structure problems and meaning."
Paper 4	T5	1	"The rewrite code was not provided, even though it was necessary for many sentences with structural and meaning-related issues." ★
Paper 4	T6	1	"AI ignored content feedback for the body paragraphs while giving detailed feedback for form and mechanical mistakes." ★

Paper 4	T7	1	"The absence of the rewriting code was a major problem on this paper, and instead of RW, some other codes were provided, but I do not believe it would help student to correct the mistake, because the problem was in meaning, or sentence structure, and this cannot be resolved without RW code."
Paper 4	T8	1	"With 9 rejected items, I thought the rejected option. I believe this paper requires major changes that would be highly time-consuming for a teacher. AI is not helping at all. Adding burden." ★
Paper 5	T1	2	"The number of partially accepted feedback codes are a lot, accepted one are also great in number while there are 3 rejected. There are still overlooked errors that should be added."
Paper 5	T2	1	"There were some problems with the outline of the feedback. Some missing lines that are not provided error description."
Paper 5	T3	1	"A LOT of overlooked errors, and inaccurate error codes." ★
Paper 5	T4	1	"There are many missing aspects like missing error descriptions and tick codes for correct sentences."
Paper 5	T5	1	"Several aspects were missing, such as error descriptions and tick codes for correct sentences."
Paper 5	T6	2	"The number of accepted and partially accepted error codes are dominant, and there are few rejected ones. There are considerable number of overlooked error codes throughout the essay."
Paper 5	T7	1	"The missing lines is one of the major problems in this paper. So many missing lines led me to think of rejection. Although presence of partial items, missing lines that are not given tick code in excel is not helpful for the student." ★
Paper 5	T8	1	"Rejected because so many missing lines, not marked tick sentences. And sentences needing RW, but did not get RW code."
Paper 6	T1	3	"Mostly correct feedback and error codes."
Paper 6	T2	3	"There are few mistakes made by AI. Overall, the feedback is useful."
Paper 6	T3	3	"Minor errors, overall accurate feedback." ★
Paper 6	T4	3	"This feedback seems sufficient to guide the teacher while giving feedback."
Paper 6	T5	3	"This feedback appears adequate to guide the teacher in providing feedback."
Paper 6	T6	3	"Most of the feedback provided is quite beneficial except for a few things."

Paper 6	T7	3	"This was really helpful feedback with many correct and useful feedback, but the paper itself does not have complex problems and is written well." ★
Paper 6	T8	3	"Very useful feedback, but there are few mistakes compared to other papers, so maybe AI found it easy to manage that paper." ★
Paper 7	T1	2	"Major overlooked critical for ss learning." ★
Paper 7	T2	2	"There are 8 reject items, 3 partial items and 8 accepted items. Because of the close numbers in reject items and accepted items, number 2 seems okay here."
Paper 7	T3	1	"A LOT of overlooked errors, and inaccurate error codes."
Paper 7	T4	2	"There are 9 reject items, 2 partial items and 8 accepted items. Thus giving number 2 seems fair for this paper."
Paper 7	T5	1	"There are 10 rejected items, 1 partially accepted item, and 8 accepted items. Therefore, assigning a score of 3 seems fair for this paper."
Paper 7	T6	1	"Major overlooked errors are critical for students' learning."
Paper 7	T7	2	"The items accepted, partials are close in number. RW global errors were missed by AI here in this paper as well." ★
Paper 7	T8	2	"Accepted items and partial items are more compared to rejected items. The major problem was overlooked errors, incorrect and redundant codes, and basic punctuation problems were missed."
Paper	Teacher	Rating	Comment
Paper 8	T1	2	"The number of accepted, (rejected ones are due to lack of tick to the correct sentences - missing lines in excel as well as major missing errors) and partial acceptances are a few."
Paper 8	T2	1	"There were so many missing lines and rejected items. AI has processing problems I guess." ★
Paper 8	T3	1	"Missing lines lead to rejection, especially lack of constructive feedback, as AI did not tick the sentences in word and excel." ★
Paper 8	T4	1	"Majority items are Rejected items are a lot due to missing line (does not mark the sentences correct with tick, or missed the lines with errors)."
Paper 8	T5	1	"RW items were a lot, but it wasn't detected by AI."
Paper 8	T6	1	"The number of accepted, partially accepted more, but rejected errors are major ones. The guidance regarding the content is insufficient."
Paper 8	T7	1	"The missing lines is one of the major problems in this paper. So many missing lines led me to think of rejection."

Paper 8	T8	1	"The accepted items and partial items were few in number, while reject items were a lot. No feedback for wrong modal usages, no feedback for wrong for subjects-(should be in plural form) or no RW feedback for sentences with expression problems." ★
Paper 9	T1	2	"The number of accepted are higher than rejected and partially rejected ones. (18 accept, 8 reject, 5 partially accepted) in total but there are 2 lines (line 14, 29) with overlooked errors which were not added to excel file created by AI." ★
Paper 9	T2	2	"A lot of overlooked codes in this paper, as well as reject items."
Paper 9	T3	2	"Sometimes important mistakes, some wrong error code placements."
Paper 9	T4	2	"13 items were accepted, 9 rejected, and 9 partially accepted. This made me think that partial acceptance would be fair. Overlooked error codes are a lot. Regarding content feedback, I really liked it and I would use it." ★
Paper 9	T5	2	"Because of the missing lines, I have changed my decision from 'accept' to 'partially accept'."
Paper 9	T6	2	"The number of errors accepted is considerably higher compared to the number of rejected and partially accepted ones. There are 16 accepted, 9 rejected due to mostly sentences without tick, and 4 partially accepted errors in total. There are two lines (line 14-29) with overlooked mistakes in the word file which were not identified and marked in the Excel file."
Paper 9	T7	2	"The missing lines is one of the major problems in this paper, (both Ticked sentences and no error description are available), but still accepted items, plus partial ones are a lot in number. Word choice problems were missed or wrong code given."
Paper 9	T8	2	"Accepted and partial ones are a lot in number, while also coupled with a few reject items. But due to reject items, I opted for partial acceptance."
Paper 10	T1	2	"Majority of lines with error codes are partially accepted. (9 accept, 8 partial, 9 reject) However, there still 8 accepted lines with error codes. There are major overlooked errors about content." ★
Paper 10	T2	2	"A lot of overlooked codes in this paper, as well as reject items. But accepted ones are also a lot."
Paper 10	T3	2	"Well, I think this last paper also have overlooked errors due to WF and RW codes missing for tone, phrasing and structural issues. Accepted and partial items are more in number while rejected items are also not few. So, I could only accepted partially." ★
Paper 10	T4	2	"14 items were rejected, 10 accepted, 3 partial items. Items were rejected because of wrong and redundant codes, overlooked codes, and due to missing lines not marked with tick code. I also partially

			accepted content feedback because body 1 and 2 have similar ideas, and this was not provided any feedback and if student add similar idea to develop it, the student would fail developing the content of the essay."
Paper 10	T5	2	"This was the paper where I have identified many wrong codes aside from the overlooked ones."
Paper 10	T6	2	"Most lines with error codes are partially accepted. (9 accept, 10 partial and 9 reject) However, there are still 9 accepted lines with error codes. There is major overlooked feedback about content."
Paper 10	T7	2	"Some good feedback descriptions like line 5, line 7. The rest of the feedback made me decide on partial acceptance."
Paper 10	T8	2	"Partial items and accepted ones were more than rejected items. I think I did not notice wrong word choices."

3. Reliability calculations of Teachers' Rubric Ratings

Paper Rating Analysis - Corrected Table

Paper	Rating (Low)	Rating (Mid)	Rating (High)	Category	Agreement Level
Paper 1 (Pilot)	0	8	0	Mid Accuracy	Full agreement (100%)
Paper 2	0	8	0	Mid Accuracy	Full agreement (100%)
Paper 3	0	8	0	Mid Accuracy	Full agreement (100%)
Paper 4	7	1	0	Low Accuracy	High agreement (87.5%)
Paper 5	5	3	0	Low Accuracy	Moderate agreement (62.5%)
Paper 6	0	0	8	High Accuracy	Full agreement (100%)
Paper 7	3	5	0	Mid Accuracy	62.5%
Paper 8	7	1	0	Low Accuracy	High agreement (87.5%)
Paper 9	0	8	0	Mid Accuracy	Full agreement (100%)
Paper 10	0	8	0	Mid Accuracy	Full agreement (100%)

Paper	Rating Split	Agreement %
Paper 5	5 Low, 3 Mid	62.5%
Paper 7	3 Low, 5 Mid	62.5%
Paper 4	7 Low, 1 Mid	87.5%
Papers 1–3, 6, 9–10	8 full agreement	100%

4.Coefficient of Variation for Rubric Ratings

Rating Category	Coefficient of Variation	Interpretation
Overall Ratings	7.89%	Very good reliability
Low Accuracy (1)	6.12%	Excellent reliability
Mid Accuracy (2)	8.34%	Very good reliability
High Accuracy (3)	0.00%	Perfect reliability

5.Qualitative Analysis Table for the Accuracy Evaluations for error-coded feedback from Interviews, and Representative Quotes (highlighted)

Sub-themes	Sub-codes	Counts	Teachers Mentionings	Quotes Notes: Highlighted representative quotes were used in the text.
Overlooked errors	General Statements	4	T1, T4, T7, T5	T1: "the major ones were the overlooked errors." T4: "I just realized many overlooked errors." T7: "overlooked codes were the major issues." T5: "most of the errors were overlooked."
	Code 1: Missing RW code	8	T1, T6, T4, T8, T3, T2, T7, T5	T1: "had to write 'rewrite' frequently because sentences either lacked meaning, were unclear, incomplete, or had some structural problems." T1: "The AI would often miss when a complete rewrite was needed.... AI couldn't identify partial

				rewritings." T2: "there were so many overlooked errors first of all, mainly RW codes..." T4: "for me, the most important one actually was rewriting sections." T8: "some sentences needed rewriting, not because of grammar mistakes." T3: "I had to add overlooked codes. use the RW code mostly for run-on sentences, incomplete ideas, repetitive structures, or tone issues." T6: "AI did not provide the RW code even when a sentence is unclear, or doesn't make sense." T5: "AI assigned individual codes when many actually needed RW codes." T7: "AI cannot provide feedback for meaning or clarity issues in a sentence. So many times, I had to provide RW code."
	Code 2: Missing auxiliary verbs	4	T1, T7, T3, T6	T1: "The most frequent corrections were missing main verbs and auxiliary verbs." T7: "I often had to add missing codes for auxiliary verbs." T3: "didn't catch missing auxiliary verbs (am, is, are, was, were) or modals." T6: "missing prepositions, auxiliary verbs, punctuation."
	Code 3: Missing main verbs	3	T1, T6, T4	T1: "The most frequent corrections were missing main verbs and auxiliary verbs." T4: "no feedback for missing verbs..." T6: "failed to identify when sentences lacked main verbs."
	Code 4: Missing punctuation	3	T7, T5, T6	T7: "Basic punctuation issues, especially for connectors like 'but' or 'so,' and they were frequently overlooked by AI." T5: "some missed punctuation errors I can remember,

				and some commas before conjunctions." T6: "missing prepositions, auxiliary verbs, punctuation."
	Code 5: Missing prepositions	2	T6, T7	T6: "missing prepositions, auxiliary verbs, punctuation." T7: "I often had to add missing codes for articles, prepositions, or missing words."
	Code 6: Missing pronouns	1	T7	T7: "I often had to add missing codes for pronouns and pronoun reference errors."
	Code 7: Parallel structure	1	T5	T5: "there was a lack of feedback on parallel structure in the lists of effects or causes, which I always have to correct manually."
	Code 8: Passive/Active structure	1	T6	T6: "It also struggled with passive and active sentence structures."
Reasons for frequent mistakes in AI error-coded feedback		6	T6, T4, T8, T2, T1 (counted based on common corrections questions)	
	Too much focus on grammar	2	T8, T6 (1x)	T6: "meaning is not its first focus, but it should be." T8: "I believe the AI focuses too much on grammar and doesn't always consider the meaning."
	Not understanding meaning issues	1	T2	T2: "AI couldn't detect that the sentence has meaning problems"
	Focusing on errors in isolation	1	T4	T4: "AI just tried to focus on them one by one, but it can be confusing."
	Performance-variation	2	T1 and T6 (1x)	T1: "when it comes to weak students, I think AI [pause] most of the

	depending on paper			feedback was problematic. Good papers, of course, they did not want so much feedback or details, so they were... they were ok." T6: "Yeah, sometimes, I could catch some accurate feedback. I think... depending on the students' paper, the dynamic of the student's paper may change. If it is a weak paper, for example, AI can be confused about it. Yeah...if I like, if I am not mistaken. This is what I have experienced."
Incorrect code application		6	T6, T8, T3, T2, T5, T7	T6: "AI provided redundant or incorrect codes, especially for word choice issues." T8: "it sometimes flagged errors for the wrong reasons." T2: "It might mark S/V, SP, or WF correctly in one sentence but miss it in another sentence." T6: "it is unclear to the AI whether the issue is word form or incorrect word choice." T3: "AI marked a word form (WF) code when it was actually a word choice (WW) problem." T3: "A student wrote 'traditional fast food restaurants' when they actually meant 'traditional local restaurants,' but AI couldn't detect the word choice mistake." T8: "Not recognizing that 'job lives' should be 'work lives'," T7: "AI cannot provide two codes at once for multiple issues in a word...one clear example was with the sentence about "growing up" in engineering. The AI provided a VF code but missed that "growing up" was also the wrong word choice." T5: "I also noticed some misidentified error codes as well (like instead of

				WW code AI gives VF code)." T5: "misidentified error codes as well (like instead of WW code AI gives VF code)." T5: "AI assigned individual codes when many actually needed RW codes instead." T3: "Failing to write codes for multiple issues in a word"

6. Qualitative Analysis of Teacher's Evaluations of AI-generated Content-feedback, and Representative Quotes (highlighted) from the teachers' interview

1.Accuracy Issues

Sub-theme	Count	Teachers Mentioning	Representative Quotes Notes: Highlighted representative quotes were used in the text.
Failure to identify repetitive ideas across paragraphs	8	T1, T6, T4, T8, T3, T2, T7, T5	T1: "The body paragraph, for example, did not give three different ideas, and AI did not see it." T6: "AI did not provide feedback when similar ideas appeared in different paragraphs." T4: "there weren't any specific comments about the identical ideas part. My students also make the same mistake all the time; they are kind of repetitive in terms of ideas. However, AI wasn't successful in flagging it, I can say" T8: "it didn't always check if paragraphs repeated similar ideas." T3: "paragraphs 1 and 2 kept repeating the same ideas...but AI ignored the repetition and didn't suggest how to fix them." T2: "AI fails to see that paragraphs have similar ideas, and it does not provide any comment on this." T7: "It should not overlook errors related to meaning and should consider context and the logical flow between sentences. This way, it could identify whether ideas are too similar, repeated, or unrelated."

			T5: "there wasn't any specific changes, I mean or specific comments about the identical ideas part."
Missing feedback for the conclusions (short-underdeveloped)	5	T1, T8, T3, T2, T5	T1: "there was a conclusion that was too short, and there was no feedback for it." T8: "should prioritize identifying real gaps in the student's writing... whether the conclusion effectively summarizes the main points or not." T3: "in conclusions, I often asked students to clearly restate the three effects or causes they had discussed instead of leaving it vague." T2: "I think it should have provided comments on the topic and the sentences of bodies, but there were no comments on that." T5: "It does not provide feedback on conclusion of essay throughout feedback although there was a need for clear restatement."
Need for more specific, actionable suggestions	4	T8, T2, T7, T5	T8: "AI gave feedback on paragraphs that didn't need specific examples. If a paragraph already made its point, there was no need to ask for details." T2: "AI should also provide revision options, like combining similar paragraphs or helping students develop more distinct supporting details." T7: "if the suggestions for body paragraphs included more specific, actionable ideas. I mean, for example, AI wrote, let's say: 'Explain more clearly how fast food chains impact local businesses.' But (inaudible) mostly students might struggle to apply feedback without um... considerate examples like suggesting 'lower prices of fast food' as an idea for development." T5: "AI always suggested adding ideas to expand body paragraphs, however it does not suggest where student can add and build on their ideas."

Misinterpretation of student meaning	2	T6, T8	T6: "AI misinterpreted this as being about marketing strategies, completely missing the student's actual points." T8: "When a student wrote about how fast food consumption created a growing sector of marketing, for example, AI misinterpreted this."
Providing direct corrections for thesis statements	1	T4	T4: "I didn't like some ideas in content feedback because sometimes AI provided direct corrections, like direct examples. For the thesis, I might want my student to write the thesis statement."

APPENDIX-15: Analysis of Interview Data for RQ3, and Representative Quotes (highlighted)

2.Pros and Cons of Refining AI feedback

Sub-theme	Count	Teachers Mentioning	Verification Quotes Notes: Highlighted representative quotes were used in the text.
Need for substantial teacher refinement (general view)	4	T8, T2 T3, T6	T8: "AI was helpful in catching basic grammar mistakes, which let me focus on meaning and coherence. But to be honest, I had to double-check its suggestions because they weren't very accurate." T2: "AI made my work harder because it missed so many errors, and I had to add them. And also, there were some wrong errors that were provided. So, I changed them a lot, and this challenged me a lot actually." T6: "Out of all the papers I reviewed, I could only accept one of them. The system has serious problems, maybe the system is not ready hocam." T3: "AI did provide a lot of feedback on basic grammar mistakes, which was useful. But... on its own, it wasn't enough."
Sub-code 1: Increased teacher time investment for refining AI feedback	4	T7, T8, T6, T2, T1	T7: "I feel like I spent a lot of time fixing AI mistakes." T8: "I had to double-check its suggestions." T6: "I spent so much time, maybe more than my usual practice, correcting them." T2: "AI made my work harder because it missed so many errors, and I had to add them." T1: "I mean for error descriptions it could be ok I think. And saving time. It makes sense, I think it saves time, but accuracy is, as I mentioned, it is still, I mean, I would still say the same thing problematic. And this has an effect on error description, like a negative effect on error description I can say."
Sub-code 2: Increased teacher cognitive load due to interpretation of AI feedback	3	T5, T1, T4	T1: "I really had difficulty while I was trying to understand some of the feedback." T4: "This difference in interpretation makes me think a lot and put in extra effort." T5: "I had to spend extra effort trying to interpret the feedback first of all because my

			interpretation and AI's interpretation don't match all the time."
Sub-code 3: Lacks in Human judgement & nuanced understanding	5	T4, T8, T3, T2, T7	T4: "I always consider whether the error codes would be helpful and understandable for the student because I have emotions, and I am a human being. I think for my student as well, but AI doesn't, I think" T8: "It struggles with nuance that a teacher can easily understand." T3: "AI focused too much on surface-level issues while missing bigger structural problems. Sometimes, a sentence didn't need small fixes. It needed a complete rewrite, for instance, but AI didn't recognize that nuance, I believe." T2: "Maybe it struggles depending on how the sentence is structured, or it doesn't interpret the student sentences well." T7: "AI needs better detection for meaning level, word choice, content, context, academic tone, and should read and analyze whether the paragraph sentences flow logically."
Easier refinement of content feedback compared to error feedback.	5	T6, T8, T2, T7, T5	T6: "To be honest, I found myself needing to revise most of the AI's feedback... even when it provided correct error codes, it often missed other errors in the same sentence or gave misleading feedback... I may prefer using content feedback, but I do not think that I will use it much." T8: "I found it more useful for content feedback." T2: "The error code part made my job harder. I think content feedback made it easier. And the feedback overall was useful for the student." T5: "For content, as well, the feedback is structured and really helpful with some small adjustments that I have made." T7: "basic grammar codes were useful as starting points. So, it effectively points out when ideas needed more specificity, like in the robots and jobs paragraph."

Appreciation of AI's guiding tone in error descriptions	1	T4	T4: "I liked AI's guiding tone sometimes. Actually, error descriptions would really be helpful if feedback was accurate because sometimes students don't understand them, like amm... the comments, I mean. And at those times, we usually provide some oral feedback to help them. However, AI could save us from this burden sometimes, I mean, if there was some accurate comment, I can say."
Promotion of Teacher self-reflection for more comprehensive feedback for papers	1	T5	T5: "umm it made me reflect more on the feedback I provide. And in fact, with AI's help, I ended up giving more comprehensive feedback than I usually do. For content, as well, the feedback is structured and really helpful with some small adjustments that I have made."

3. Usability of AI feedback and Teacher Attitudes

Sub-theme	Count	Teachers Mentioning	Verification Quotes Notes: Highlighted representative quotes were used in the text.
Reluctance to use AI feedback (mainly for error feedback)	8	T1, T6, T4, T8, T3, T2, T7, T5	T1: "I don't feel like using it. The reason for not using it is its focus on... I really personally give importance to giving content, but when I looked at the feedback, most of the content feedback was either missing or overlooked." T6: "I can say that I am not positive about it." T4: "To be honest, I think not now because I need to double-check everything. I needed to check whether the content feedback is correct or whether the error codes are correct. So, this seems a bit complicated for me right now... Content is better, I guess, but still not enough." T8: "I would use it definitely, but not for error feedback. Actually, I found it more useful for content feedback, though I'd say, I still add my own refinements for missing elements" T3: "to be honest, not as it is. Writing isn't just about correct sentences. It's about clear, logical communication, I believe. So, AI still requires too much extra effort to refine its feedback, so... I wouldn't rely on it." T2: "maybe only for providing content feedback I can use it because accuracy was

			better than the error codes feedback, let's say. But it would be again with my refinements and additions. However, I definitely think it can help me in class. Umm... as I said, for error, I can say no, for now." T7: "I would use it for, let's say, content feedback, and while adding my comments for content development and organization. Yes... good, yeah... it provides good things, but it needs teachers to look at it before giving feedback." T5: "in terms of accuracy issues, I wouldn't use it for error feedback, but I would only use it for content feedback maybe..."
Need for contextual/tailored AI training	8	T1, T6, T4, T8, T3, T2, T7, T5	T1: "there is a serious need for AI training, right?" T6: "We need a system that can better understand context, our students' context, to provide more accurate error feedback." T4: "Maybe AI needs more training on how to read student papers because it focuses too much on errors sometimes and risks overlooking meaning." T8: "needs better detection for meaning level, word choice, content, context, academic tone." T3: "AI needs to analyze writing...the way we teach it." T5: "having some technical support maybe would really help me here." T2: "AI training for these problematic points would be absolutely necessary." T7: "think the system needs several improvements."
Need for university-level AI training support	2	T1, T5	T1: "think instead of individual efforts, the university can help us with this kind of AI training." T5: "if there was a unit at the university that could work on improving the accuracy of AI feedback, specifically to fit it to our teaching practices."

4.Improvement Suggestions

Sub-theme	Count	Teachers Mentioning	Quotes Notes: Highlighted representative quotes were used in the text.
Prioritize meaning over grammar	6	T4, T8, T3, T2, T5, T7	T4: "Sometimes meaning is more important, but if we focus directly on mistakes, we risk everything." T8: "If there's no issue with meaning, then it can move on to basic grammar corrections." T2: "AI should check meaning and context first before grammar." T5: "Like we do, AI should check meaning first, not grammar." T3: "AI needs to go beyond grammar and assess effective communication ...Fixing grammar won't help if the sentence doesn't make sense." T7: "The most important improvement should be within meaning, context, and logical sentence flow over grammar."
Better detection of context problems and solutions	4	T1, T6, T8, T5	T1: "if there were some extra details about the context and logical flow between sentences." T6: "should consider context and the logical flow between sentences." T8: "It should not see a sentence in isolation but always consider each sentence in relation to the one before and after it." T5: "Given our student profile and context, I think, it is unlikely that students have the background knowledge to add statistics because that would require prior research."
Better detection of repetitive content & solutions	4	T8, T3, T2, T7	T8: "This way, it could identify whether ideas are too similar, repeated, or unrelated." T3: "It should first include some improvements on identifying logical fallacies and tracking repetitions." T2: "If it detects similar ideas, it should give students the option to either write something different or combine the paragraphs." T3: "circular reasoning, for example, track repetition across paragraphs, and detect underdeveloped arguments." T2: "AI should compare body paragraphs to check for repetitive content." T7: "It should not overlook errors related to meaning and should consider context and the logical flow between sentences. This way, it could identify

			whether ideas are too similar, repeated, or unrelated."
Better detection of missing feedback and solution	1	T6	T6: "AI needs to identify when ideas are repetitive across paragraphs and provide feedback on all aspects of the essay, including the conclusion ... If the word limit is a concern for AI, it should not be restricted"
Solutions for providing specific/ actionable ideas for content	1	T2	T2: "usually quote their sentences and ask them to explain. Or if the idea needs expansion, I add a note and say, like, 'You need to add ideas, give examples' kind of thing for this idea."

5.AI Familiarity

Experience Level	Count	Teachers Mentioning	Quotes
Users having no experience for using AI for feedback	3	T1, T2, T4	T1: "Never ever tried. I have never used AI to give feedback." T2: "I have used ChatGPT for just (...) lesson plans or materials. Other than that, no. For feedback, also no." T4: "I used ChatGPT a lot, but for various purposes. Most of the time, just to get some ideas for lesson planning and some activities. However, I don't really use it for feedback."
Limited experience with AI for feedback	4	T6, T7, T8, T5	T6: "I used it once for feedback for an essay. But I didn't use it much." T7: "I just only know AI, ChatGPT as an AI. I did not use it for the whole feedback because, you know, we have to be sure about it." T8: "I have only used it once in class. I provided the prompt to a student, and then they took a picture of their writing and wrote my prompt to ChatGPT, and then they got feedback. And it was actually very accurate, but it was only A1 level. So [Pause] it was... it was specifically concentrated on the grammar errors, so I do not know if it is accurate in umm... organization, content-related things, errors, and issues as well." T5: "I haven't used them, not much let's say. So maybe just to get an idea how to give feedback on student paper, but to be honest, I did not get any imm.. help from AI to get feedback for my students."
More experienced AI users	1	T3	T3: "I can say I am very much familiar with these AI platforms. I've been using them in both my own research and also, in my writing lessons, I've made use of ChatGPT for giving feedback."

APPENDIX-16: Ethics Committee Approval form

Evrak Kayıt Tarihi: 08.10.2024	Protokol No: 796639	Tarih: 31.10.2024



ANADOLU ÜNİVERSİTESİ
SOSYAL VE BEŞERÎ BİLİMLER BİLİMSEL ARAŞTIRMA VE YAYIN ETİĞİ KURULU
KARAR BELGESİ

ÇALIŞMANIN TÜRÜ:	Doktora Tez Çalışması
KONU:	Eğitim Bilimleri
BAŞLIK:	Geribildirim Geleceği: Öğretmenlerin Yapay Zekâ Geribildirim Doğruluğuna İlişkin Algılarını, Geliştirme Uygulamalarını ve Yapay Zekâ Tarafından Oluşturulan Geribildirim Sürecine İlişkin Görüşlerini Keşfetmek
PROJE/TEZ YÜRÜTÜCÜSÜ:	Prof. Dr. Ali MERÇ
TEZ YAZARI:	Burcu GÜNDOĞDU
ALT KOMİSYON GÖRÜŞÜ:	-
KARAR:	Olumlu

APPENDIX-17: Sample Consent form

ARAŞTIRMA GÖNÜLLÜ KATILIM FORMU

(Öğretmen için)

Bu çalışma, “Geribildirim Geleceği: Öğretmenlerin Yapay Zekâ Geribildirim Doğruluğuna İlişkin Algılarını, Geliştirme Uygulamalarını ve Yapay Zekâ Tarafından Oluşturulan Geribildirim Sürecine İlişkin Görüşlerini Keşfetmek” başlıklı bir araştırma çalışmasıdır. Çalışma, Burcu GÜNDOĞDU tarafından yürütülmekte olup, Yabancı Diller bölümü hazırlık biriminde öğretim görevlisi olan öğretmenlerin AI tarafından üretilen geribildirimleri kendi içgörülerıyla nasıl birleştirdiği, öğretmenlerin geribildirimleri gözden geçirirken yaptıkları geliştirmeleri / önerileri, AI geribildirimlerinin doğruluğuna dair değerlendirmeleri ve bu deneyimlere dayanarak, AI geribildirimlerinin gelecekte sınıf içinde nasıl kullanılabileceği hakkındaki görüşlerine ışık tutmayı amaçlamaktadır. Bu çalışmaya katılımınız tamamen gönüllülük esasına dayanmaktadır.

- Çalışmanın amacı doğrultusunda, sizlere Yapay Zekâ geribildiriminin doğruluğunu gözden geçirme ve geliştirme konusunda kullanılacak araçlar hakkında bir eğitim verilecektir. Ayrıca, AI geribildiriminin doğruluğunu dereceli puanlama aracı (rubrik) kullanarak nasıl ölçeceğinize dair bir eğitim verilecektir. Araştırma için, sizlerin AI geribildirimi gözden geçirme ve entegrasyonu amacıyla yaptığınız geliştirmeleri Word dokümanına işlemeniz ve bu değişiklikleri Google Excel tablosunda raporlamanız istenecektir. Burada, AI geribildirimlerinde neleri kabul ettiğinizi, etmediğinizi veya kısmen kabul ettiğinizi belirtmeniz ve nedenlerini açılır listeden seçmeniz istenerek veri toplanacaktır. Sonrasında, sizlerden AI geribildiriminin doğruluğunu rubrik ile değerlendirmeniz istenerek veri toplanacaktır. Son olarak, bireysel görüşmeler yapılarak bu süreçte yaşadığınız avantajlar ve dezavantajlar, geribildirim doğruluğuna dair görüşleriniz ve gelecekte sınıfta kullanıma yönelik önerileriniz veri olarak toplanacaktır.
- Bu doğrultuda, çalışmaya katılımcı olarak yer almanız için izniniz gerekmektedir.
- Öğretmenlerin isimlerini yazmaları veya kimliklerini açığa çıkaracak herhangi bir bilgi vermeleri zorunlu değildir; araştırmada katılımcıların isimleri gizli tutulacaktır.
- Araştırma kapsamında toplanan veriler yalnızca bilimsel amaçlar doğrultusunda kullanılacak, araştırmanın amacı dışında veya başka bir araştırmada kullanılmayacak ve gerekmesi halinde sizin (yazılı) izniniz olmadan başkalarıyla paylaşılmayacaktır.
- Talep etmeniz halinde, toplanan verileri inceleme hakkınız bulunmaktadır.
- Öğretmenlerden toplanan veriler, araştırma süresi boyunca korunacak ve araştırma bitiminde arşivlenecek veya imha edilecektir.
- Çalışma, öğretmenlerin gözden geçirme ve değerlendirme süreçlerini verimli şekilde yapabilmeleri için uygun zamanlamayla planlanacaktır.
- Öğretmenler, katılım sırasında herhangi bir sebepten rahatsızlık hissederseniz çalışmadan istedikleri zamanda ayrılabilirler. Çalışmadan ayrılmalari durumunda, öğretmenlerden toplanan veriler çalışmadan çıkarılacak ve imha edilecektir.

Gönüllü katılım formunu okumak ve değerlendirmek üzere ayırdığınız zaman için teşekkür ederim. Çalışma hakkındaki sorularınızı Antalya Bilim Üniversitesi, Yabancı Diller bölümünden Burcu Gündoğdu'ya yöneltebilirsiniz.

Araştırmacı Adı: Burcu GÜNDOĞDU

Adres:

İş Tel:

Cep Tel:

Bu çalışmanın dersim kapsamında sınıfta uygulanmasına onay veriyorum ve araştırma sonucunda elde edilen verilerin bilimsel amaçlarla kullanılmasını kabul ediyorum.

(Lütfen bu formu doldurup imzaladıktan sonra veri toplayan kişiye veriniz.)

Katılımcı Ad ve Soyadı:

İmza:

ÖZGEÇMİŞ

Adı Soyadı: Burcu GÜNDOĞDU

Yabancı Dil: İngilizce

Eğitim Bilgileri:

Doktora: 2018–, Anadolu Üniversitesi, Lisansüstü Eğitim Enstitüsü, Yabancı Diller Eğitimi Anabilim Dalı, İngilizce Eğitimi Programı, Eskişehir.

Yüksek Lisans: 2013–2016, İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, İngiliz Dili ve Edebiyatı (Tezli), İstanbul.

Lisans: 2009–2012, İstanbul Kültür Üniversitesi, Fen-Edebiyat Fakültesi, İngiliz Dili ve Edebiyatı Bölümü, İstanbul.

Meslek Denevimi:

2022–, devam ediyor: Öğretim Görevlisi, Antalya Bilim Üniversitesi, Yabancı Diller Yüksek Okulu. Güllük Yerleşkesi, Antalya

2015–2022: Araştırma Görevlisi, İstanbul Kültür Üniversitesi, İngilizce Öğretmenliği Programı, Yabancı Diller Eğitimi Bölümü

Yayınları ve/veya Bilimsel/Sanatsal Faaliyetleri:

Gündoğdu, B. & Merç, A. (2022). A Systematic Review of Tech-supported Collaborative Creativity Practices in the Field of Education. *Journal of Learning and Teaching in Digital Age*, 7(1), 76–89. <https://doi.org/10.53850/joltida.953760>

Gündoğdu, B. & Alkayalar, A. A. (2021). Early Teacher Identity and Initial Teaching Beliefs of EFL Pre-service Teachers During Covid-19 Pandemic: What Changed? *Eurasian Journal of Teacher Education*, 2(3), 195–220.

Uğur Usta, D. & Gündoğdu, B. (2019). Impact of Reading Strategy Use on Reading Motivation in Turkish EFL Context: A Mixed Method Study. *The Literacy Trek*, 4(2), 15–30.

Gündoğdu, B. (2018). *A Contemporary Outlook on Gender, Identity, Violence and Melancholia*. Türkiye Alim Kitapları. ISBN: 978-3639812466

Gündoğdu, B. (2016). Violence and Melancholia: The Story of the Melancholic Butterfly Generation. *International Journal of Social Sciences and Education Research*, 2(2). <https://doi.org/10.24289/ijsser.279049>

Gündoğdu, B. (2015). Resistance to Sexist Medical Community with Semiotic Language in Augustine (Big Hysteria). *Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, (20), 1–9.

Gündoğdu, B. (2015). Towards Remembering the Sufi Philosophy: Deconstruction of Boundary Oppositions in Doris Lessing's Briefing for a Descent into Hell. *Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, (21), 1–9.
<https://doi.org/10.5505/pausbed.2015.44154>

Ödülleri:

2019, –, devam ediyor: TÜBİTAK 2211-A Genel Yurt İçi Doktora Burs Programı, Ankara.

