

REPUBLIC OF TÜRKİYE
SAKARYA UNIVERSITY
INSTITUTE OF EDUCATIONAL SCIENCES
DEPARTMENT OF FOREIGN LANGUAGE EDUCATION
ENGLISH LANGUAGE EDUCATION PROGRAM

**AI-ASSISTED ASSESSMENT OF ESL WRITING: UNIVERSITY
INSTRUCTOR PERCEPTIONS & HUMAN-AI SCORING CORRELATIONS**

A MASTER'S THESIS

BERKAY YİĞİT

SUPERVISOR
ASSIST. PROF. DR. ALI İLYA

JUNE 2025

REPUBLIC OF TÜRKİYE
SAKARYA UNIVERSITY
INSTITUTE OF EDUCATIONAL SCIENCES
DEPARTMENT OF FOREIGN LANGUAGE EDUCATION
ENGLISH LANGUAGE EDUCATION PROGRAM

**AI-ASSISTED ASSESSMENT OF ESL WRITING: UNIVERSITY
INSTRUCTOR PERCEPTIONS & HUMAN-AI SCORING CORRELATIONS**

A MASTER'S THESIS

BERKAY YİĞİT

SUPERVISOR
ASSIST. PROF. DR. ALI İLYA

JUNE 2025

DECLARATION

In this study, which I have prepared in accordance with the Thesis-Project Writing Guide of the Sakarya University Institute of Educational Sciences, I hereby declare that:

- All information and documents included in the thesis have been obtained and presented within the framework of academic and ethical principles,
- I have cited and referenced all sources I have made use of,
- I have not made any alterations to the data used,
- I have not submitted this thesis, in whole or in part, as another thesis work.



Berkay YİĞİT

SIGNATURES PAGE

This thesis entitled “AI-Assisted Assessment of ESL Writing: University Instructor Perceptions & Human-AI Scoring Correlations”, prepared by Berkay YİĞİT, was accepted on 27/06/2025 by the unanimous vote of the following jury as a Master’s thesis in the Department of Foreign Language Education, English Language Education Program, Institute of Educational Sciences, Sakarya University.

Thesis Jury

Chairperson Prof. Dr. Ayşegül DALOĞLU

Member (Supervisor) Assist. Prof. Dr. Ali İLYA

Member Assoc. Prof. Dr. Gülden KAYA UYANIK

ACKNOWLEDGEMENTS

I would like to dedicate this thesis to Bıdık, who, despite being thoroughly annoyed with me for being constantly busy and having the audacity to leave the house for errands, has remained loyally by my side throughout this journey. He has been my wonderfully grumpy little companion through every step of this process, and I'm truly grateful to have him with me.

I am deeply thankful to my family—my mum Hülya, my dad Selahattin, and my brother Sercan—for creating such a calm, supportive, and wonderfully stress-free environment while I wrestled with this thesis. Their quiet but unwavering support made all the difference in the world.

I'm grateful to have had Assist. Prof. Dr. Ali İLYA as my thesis advisor. His supportive guidance and insights were helpful throughout this journey. In particular, his willingness to involve me in multiple TÜBİTAK projects offered invaluable exposure to the dynamics of project design, coordination, and academic collaboration in Turkey. He not only enriched the development of this thesis but also contributed to my academic growth.

I also want to give special thanks to Assoc. Prof. Dr. Gülden KAYA UYANIK, whose infectious positive attitude and genuinely engaging classes made learning such a joy. Her motivational approach consistently lifted my spirits, and her insightful guidance helped shape not just this thesis but my entire academic perspective. Her wisdom and encouragement have been truly invaluable.

Together, these remarkable academics have been the cornerstone figures who made this thesis possible, and I feel incredibly fortunate to have learned from both of them.

*“In the garden of academia, every page written is a petal grown—
with thorns, sun, and unexpected rain.”*

ABSTRACT

AI-ASSISTED ASSESSMENT OF ESL WRITING: UNIVERSITY INSTRUCTOR PERCEPTIONS & HUMAN-AI SCORING CORRELATIONS

Berkay YİĞİT, Master's Thesis

Supervisor: Assist. Prof. Dr. Ali İLYA

Sakarya University, 2025

This thesis investigates the integration of artificial intelligence (AI) into English writing assessment through two interconnected studies: a survey of university instructors' perspectives and a correlational analysis of human and AI scoring of argumentative English as a Second Language (ESL) essays. First, a cross-sectional survey ($N = 96$) captured instructors' usage patterns, trust, perceived effectiveness, ethical concerns, and future outlook regarding AI tools. Usage of general purpose AI tools such as ChatGPT and Grammarly were prevalent. The survey additionally revealed high adoption (82%), moderate confidence in AI feedback for grammar and organization, but considerable reservations about AI's grading capacity, ethical implications, and the necessity of sustained human oversight. Second, a correlational design examined scores assigned by 4 human raters and 25 AI tools on 60 argumentative ESL essays, employing Krippendorff's alpha (K_α), ICC_{2k} , paired t-tests, Wilcoxon signed-rank tests, and Pearson (P_r) and Spearman (S_ρ) correlation coefficients. Human raters demonstrated excellent inter-rater reliability ($K_\alpha > 0.7$; $ICC_{2k} > 0.94$), while AI tools showed moderate inter-rater reliability ($K_\alpha \leq 0.5$; $ICC_{2k} > 0.94$). AI scores correlated moderately to strongly with human means ($P_r \in [0.51, 0.65]$; $S_\rho \in [0.52, 0.64]$) but systematically under-scored by approximately 0.5 to 1 points on all criteria and 2.5 points on total score ($p < 0.001$; Cohen's $d \in [-0.98, -0.47]$). Among AI tools, OpenAI o3-mini assigned scores which consistently achieved strong correlation with human scores. These findings suggest that, although contemporary AI tools can approximate human judgment in writing assessment, it exhibits systematic bias and limited construct validity. The study underscores the importance

of human and AI collaboration, ethical safeguards, and ongoing validation when deploying AI for high-stakes language assessment, pointing to a complementary role for AI that enhances efficiency without displacing human oversight.

Keywords: Artificial Intelligence, Large Language Models, Assessment and Evaluation, Educational Technology, ESL Writing, Perceptions of University Instructors



ÖZET

YAPAY ZEKA DESTEĞİ İLE İNGİLİZCE YAZMA DEĞERLENDİRMESİ: ÜNİVERSİTE ÖĞRETİM ELEMANLARININ GÖRÜŞLERİ VE İNSAN-YZ PUANLAMA KORELASYONLARI

Berkay YİĞİT, Yüksek Lisans Tezi

Danışman: Dr. Öğr. Üyesi Ali İLYA

Sakarya Üniversitesi, 2025

Bu tez, İngilizce yazma değerlendirmelerine yapay zeka (YZ) araçlarının entegrasyonunu, üniversite öğretim elemanlarının görüşlerini ölçen bir anket çalışması ve insan ile YZ puanlamalarının ilişkisel analizi olmak üzere iki birbirini tamamlayan çalışma aracılığıyla ele almaktadır. İlk olarak, kesitsel bir anket ile ($N = 96$) ChatGPT ve Grammarly gibi genel amaçlı YZ araçlarının yaygın kullanımını, bu araçlara yönelik güven düzeylerini, algılanan etkinliklerini, etik kaygılarını ve gelecek beklentilerini ortaya koymuştur. Sonuçlarda %82'lik yüksek bir benimseme oranı, dilbilgisi ve organizasyon geri bildirimine ilişkin orta düzeyde güven, ancak puanlama kapasitesi, etik boyutlar ve insan denetiminin gerekliliği hususunda belirgin çekinceler olduğu gözlemlenmiştir. İkinci olarak, 4 insan değerlendirici ile 25 YZ araçlarının İngilizce (ikinci dil olarak, ESL) kullanılarak yazılan 60 tartışmacı kompozisyonlara verdikleri puanlar Krippendorff α , ICC_{2k} , eşleştirilmiş t-testler, Wilcoxon işaretli sıralı testleri ile Pearson (P_r) ve Spearman (S_ρ) korelasyon katsayıları kullanılarak incelenmiştir. İnsan değerlendiriciler mükemmel değerlendiriciler-arası güvenirlilik ($\alpha > 0,7$; $ICC_{2k} > 0,94$), ancak YZ araçları orta düzeyde değerlendiriciler-arası güvenirlilik sergilemiştir ($\alpha \leq 0,5$; $ICC_{2k} > 0,94$). YZ puanları, insan ortalamaları ile orta ila güçlü korelasyon göstermiştir ($P_r \in [0,51; 0,65]$; $S_\rho \in [0,52; 0,64]$), ancak analitik kriterlerde ortalama 0,5, toplam puanda ortalama 2,5 puan düşük vermiştir ($p < 0.001$; Cohen $d \in [0,48; 0,98]$) ve bu araçlar arasında OpenAI o3-mini en yüksek P_r değerlerini elde etmiştir. Bu bulgular, güncel YZ araçlarının yazılı değerlendirmede insan kararlarına

yaklaşabildiğini ancak sistematik yanılma ve sınırlı yapı geçerliliği taşıdığını ortaya koyarken, yüksek riskli dil değerlendirmelerinde YZ'nin insan denetimini tamamlayıcı, etik önlemler ve sürekli doğrulama gerektiren bir rol üstlenmesi gerekliliğini vurgulamaktadır.

Anahtar Kelimeler: Yapay Zeka, Büyük Dil Modelleri, Ölçme ve Değerlendirme, Eğitim Teknolojisi, İkinci Dil Olarak İngilizce'de Yazma, Üniversite Öğretim Elemanlarının Algıları



TABLE OF CONTENTS

DECLARATION	i
SIGNATURES PAGE	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
ÖZET	vi
LIST OF TABLES	xiii
LIST OF FIGURES	xv
LIST OF ACRONYMS	xvi
CHAPTER I. INTRODUCTION	1
1.1. Background to the thesis	1
1.2. Statement of the problem	3
1.3. The aim of the thesis	3
1.4. Research questions	4
1.5. Limitations of the thesis	4
1.6. List of definitions	5
CHAPTER II. LITERATURE REVIEW	7
2.1. A historical perspective into automated essay scoring	7
2.1.1. Early developments and integration of natural language processing techniques	7
2.1.2. Contemporary automated essay scoring: Powered by artificial intelligence .	8
2.1.3. Challenges, critiques, and future directions	9
2.1.4. Automated writing evaluation and its relationship to automated essay scoring	10
2.2. The beginning of large language models	12
2.2.1. The transformer deep learning architecture	13

2.2.2.	Generative pre-trained transformers	13
2.2.3.	Zero-shot in large language models	14
2.2.4.	Large language models' emergence in education	14
2.3.	Use of artificial intelligence in assessment: English writing context	15
2.3.1.	Methodological frameworks in writing assessment	15
2.3.2.	Human and artificial intelligence grading	17
2.3.3.	Educators' perceptions of artificial intelligence in assessment	18
CHAPTER III. METHODOLOGY		20
3.1.	Research design	20
3.2.	Participants	21
3.3.	Data collection instruments	25
3.4.	Data collection procedure	27
3.5.	Data preparation and cleaning for analysis	29
3.6.	Data analysis procedures	30
CHAPTER IV. FINDINGS		32
4.1.	University instructors' view on use of artificial intelligence in English writing assessment	32
4.1.1.	Artificial intelligence adoption patterns	32
4.1.2.	Artificial intelligence employment strategies	34
4.1.3.	Perceptions of effectiveness of artificial intelligence	36
4.1.4.	Self-reported confidence, fairness, and pedagogical alignment of artificial intelligence	40
4.1.5.	Perceived alignment with personal criteria and rubric-based assessment	42
4.1.6.	Perceptions of ethical implications of using artificial intelligence	44
4.1.7.	Role of artificial intelligence in the future	45

4.1.8.	Human oversight in artificial intelligence	48
4.2.	Human and artificial intelligence scores for argumentative ESL essays . . .	49
4.2.1.	Human scores analysis	49
4.2.1.1.	IELTS-WT2B criterion: Coherence and cohesion	49
4.2.1.2.	IELTS-WT2B criterion: Grammatical range and accuracy	50
4.2.1.3.	IELTS-WT2B criterion: Lexical resource	51
4.2.1.4.	IELTS-WT2B criterion: Task response	52
4.2.1.5.	IELTS-WT2B criteria: The total score	53
4.2.2.	Artificial intelligence scores analysis	54
4.2.2.1.	IELTS-WT2B criterion: Coherence and cohesion	54
4.2.2.2.	IELTS-WT2B criterion: Grammatical range and accuracy	56
4.2.2.3.	IELTS-WT2B criterion: Lexical resource	57
4.2.2.4.	IELTS-WT2B criterion: Task response	58
4.2.2.5.	IELTS-WT2B criteria: The total score	59
4.2.3.	Drawing comparisons between human and artificial intelligence scores . . .	60
4.2.3.1.	IELTS-WT2B criterion: Coherence and cohesion	62
4.2.3.2.	IELTS-WT2B criterion: Grammatical range and accuracy	64
4.2.3.3.	IELTS-WT2B criterion: Lexical resource	65
4.2.3.4.	IELTS-WT2B criterion: Task Response	65
4.2.3.5.	IELTS-WT2B criteria: The total score	66
4.2.3.6.	Top 5 artificial intelligence raters by correlation coefficients	67

CHAPTER V. CONCLUSION, DISCUSSION, AND SUGGESTIONS	70
5.1. Discussion	70
5.1.1. Adoption patterns across contexts	70
5.1.1.1. Historical trajectory and the scope of current usage	70
5.1.1.2. Frequency and functions of use	70
5.1.1.3. Perceived effectiveness and trust	71
5.1.2. Pedagogical applications and the synergy between human and artificial intelligence	71
5.1.2.1. Hybrid models in practice	71
5.1.2.2. Criterion-specific insights	71
5.1.2.3. Adoption trajectory and future outlook	72
5.1.2.4. Implications for practice	72
5.1.3. Divergences in perceived versus actual effectiveness	73
5.1.3.1. Surface-level features: Grammar and syntax	73
5.1.3.2. Higher-order skills: Content, organization, and task response	73
5.1.3.3. Overall scoring: Consistency versus accuracy	74
5.1.3.4. Implications for practice and further research	74
5.1.4. Contextual factors in the performance of artificial intelligence raters	74
5.1.5. Ethical and pedagogical implications	75
5.2. Conclusion and Suggestions	76
5.2.1. Key findings	76
5.2.2. Interpretation and broader implications	77
5.2.3. Directions for future research	78
5.2.4. Summary	78
REFERENCES	80

APPENDICES	97
APPENDIX A.Configuration of large language models	97
1. System prompt	97
2. User prompt	104
3. Response format	105



LIST OF TABLES

3.1	List of AI Tools Used as AI Raters	23
4.1	Frequencies and Percentages for Survey Item 5	32
4.2	Frequencies and Percentages for Survey Item 6	33
4.3	Frequencies and Percentages for Survey Item 7	35
4.4	Frequencies and Percentages for Survey Item 8	35
4.5	Frequencies and Percentages for Survey Item 9	37
4.6	Frequencies and Percentages for Survey Item 10	38
4.7	Frequencies and Percentages for Survey Item 11	38
4.8	Frequencies and Percentages for Survey Item 12	39
4.9	Frequencies and Percentages for Survey Item 13	40
4.10	Frequencies and Percentages for Survey Item 14	41
4.11	Frequencies and Percentages for Survey Item 15	41
4.12	Frequencies and Percentages for Survey Item 30	42
4.13	Frequencies and Percentages for Survey Item 16	43
4.14	Frequencies and Percentages for Survey Item 17	43
4.15	Frequencies and Percentages for Survey Item 29	45
4.16	Frequencies and Percentages for Survey Item 32	46
4.17	Frequencies and Percentages for Survey Item 33	46
4.18	Frequencies and Percentages for Survey Item 35	47
4.19	Frequencies and Percentages for Survey Item 34	48
4.20	Descriptives of Aggregated Human Scores for the IELTS-WT2B “Coherence and Cohesion” Criterion	50
4.21	Descriptives of Aggregated Human Scores for the IELTS-WT2B “Grammatical Range and Accuracy” Criterion	51
4.22	Descriptives of Aggregated Human Scores for the IELTS-WT2B “Lexical Resource” Criterion	52
4.23	Descriptives of Aggregated Human Scores for the IELTS-WT2B “Task Response” Criterion	53
4.24	Descriptives of Aggregated Human Scores for the sum of IELTS-WT2B Criteria	54

4.25 Descriptives of Aggregated AI Scores for the IELTS-WT2B “Coherence and Cohesion” Criterion	55
4.26 Descriptives of Aggregated AI Scores for the IELTS-WT2B “Grammatical Range and Accuracy” Criterion	56
4.27 Descriptives of Aggregated AI Scores for the IELTS-WT2B “Lexical Resource” Criterion	57
4.28 Descriptives of Aggregated AI Scores for the IELTS-WT2B “Task Response” Criterion	58
4.29 Descriptives of Aggregated AI Scores for the Sum of IELTS-WT2B Criteria . .	59
4.30 Shapiro-Wilk Normality Test Results for Paired Difference Scores Across the IELTS-WT2B Criteria	60
4.31 Correlations Between Aggregated Human and AI Scores	61
4.32 Top 5 AI Raters by Pearson’s Correlation Coefficients	68

LIST OF FIGURES

3.1	Distribution of Participants' Academic Experience by Age Group	22
3.2	Distribution of Participants' Academic Ranks	22
4.1	Scatter Plot Comparing Mean Human and AI "Coherence and Cohesion" Scores	63
4.2	Scatter Plot Comparing Mean Human and AI "Grammatical Range and Accuracy" Scores	63
4.3	Scatter Plot Comparing Mean Human and AI "Lexical Resource" Scores	63
4.4	Scatter Plot Comparing Mean Human and AI "Task Response" Scores	63
4.5	Scatter Plot Comparing Mean Human and AI Total Scores	63



LIST OF ACRONYMS

ICC_{2k}	Intraclass Correlation Coefficient (two-way, random effects, absolute agreement model; as defined in Shrout & Fleiss, 1979)
C_α	Cronbach's Alpha (as defined in Cronbach, 1951)
C_d	Cohen's d (as defined in Cohen, 2013)
K_α	Krippendorff's Alpha (as defined in Krippendorff, 2019)
P_r	Pearson's Correlation Coefficient (as defined in Pearson, 1895)
S_ρ	Spearman's Rank Correlation Coefficient (as defined in Spearman, 1904)
AES	Automated Essay Scoring
AI	Artificial Intelligence
API	Application Programming Interface
AWE	Automated Writing Evaluation
B.A.	Bachelor of Arts
BERT	Bidirectional Encoder Representations From Transformers
CC	Coherence and Cohesion (a criterion in the IELTS Writing Task 2 Bands)
CEFR	Common European Framework of Reference for Languages (as defined in Council of Europe, 2020)
CI	Confidence Interval
EFL	English as a Foreign Language
ELT	English Language Teaching
ESL	English as a Second Language
ETS	Educational Testing Service
GPT	Generative Pre-trained Transformer
GRA	Grammatical Range and Accuracy (a criterion in the IELTS Writing Task 2 Bands)

(Continued)

GRE	Graduate Record Examinations (from Educational Testing Service)
IEA	Intelligent Essay Assessor (from Pearson Education Inc.)
IELTS	International English Language Testing System
IELTS-WT2B	IELTS Writing Task 2 Bands
LLM	Large Language Model
LR	Lexical Resource (a criterion in the IELTS Writing Task 2 Bands)
LSA	Latent Semantic Analysis
LSTM	Long Short-term Memory
M.A.	Master of Arts
ML	Machine Learning
NLP	Natural Language Processing
PEG	Project Essay Grade
RNN	Recurrent Neural Network
TOEFL	Test of English as a Foreign Language
TR	Task Response (a criterion in the IELTS Writing Task 2 Bands)

CHAPTER I

INTRODUCTION

Artificial intelligence (AI) has been infiltrating various sectors in the recent years, including education (e.g., Cumming & McDougall, 2000; du Boulay, 2016; Holmes et al., 2019; Kemal & Liman-Kaban, 2024; Khan et al., 2024). The process has been fast-forwarded thanks to developments in natural language processing (NLP) and machine learning (ML); AI tools are sure to become integrated assistants in educational settings. The use of such tools in assessing writing has already attracted significant interest (e.g., J. C. Burstein, 2003; Edwin Welch & Frick, 1993; S. Elliot, 2003; Foltz et al., 2013; Landauer et al., 1998, 2002; Page, 1966, 1968; Rudner & Liang, 2002; Shermis, 2014; Shermis & Burstein, 2003).

This chapter is an introduction to the two parallel investigations building this thesis. Together, these investigations first explore the role, adoption status, and use frequency of AI in writing assessment through the perceptions and outlook of university instructors. Later, scores assigned by human and AI raters in a writing assessment task are examined for correlations between the outputs. Through these two parallels, this thesis hopes to shed light onto multiple facets of AI: adoption by educators and accuracy by metrics.

The sections of the chapter are aligned bottom-to-top. First, a background to the research is given in hopes to enlighten the reader on why the researcher elected to investigate these topics. Secondly, the statement of the problem is described to provide a contextual framework. The aims and questions of the thesis follows the two; with assumed limitations of the research clearly listed afterward. A short list of definitions, for the terms used across the thesis, are included at the end.

1.1. Background to the thesis

The use of AI in assessment—specifically that of writing—has been gaining attention in the current decade (e.g., see Pedro et al., 2019; Stevenson & Phakiti, 2019). The possibilities offered by AI tools, as described by many of their authors, seem divine at first sight: “[The model o3] sets a new standard for math, science, coding, and visual

reasoning tasks.” (OpenAI, 2025); “[Claude 3 Opus features] top-level intelligence, fluency, and understanding.” (Anthropic, 2025); “[Gemini 2.5 Pro is] designed to tackle increasingly complex problems.” (Google, 2025); “In fact, we are able to detect the presence of AI writing with confidence.” (Turnitin, 2023) and so on. One would wonder the potential of such tools in educational contexts, such as immediate yet consistent feedback for students and grounded and systemic approach to repetitive tasks. Both of these two examples constitute the very foundation of assessing writing. The need for an investigation into the possible benefits and drawbacks of AI tools within the education context is apparent.

There already has been some exploration into the aforementioned aspects. For instance, a study comparing human and AI grading in higher education contexts has indicated that while AI tools can approximate human grading standards, discrepancies persist in scoring and feedback quality (Flodén, 2024). A different study by Lundgren (2024) found that while GPT-4 (from OpenAI) aligned with human grading standards on mean scores, it “is risk averse, seemingly avoiding assigning low or high grades” (p. 3), suggesting that AI primarily assesses “generic essay characteristics” rather than “adapting to nuanced grading criteria” (p. 4). Interestingly so, another study found that argumentative essays generated by AI were linguistically higher in quality compared to those written by humans (Herbold et al., 2023). The ability of AI tools to generate content meeting academic standards, yet falling short of analyzing similar content systematically not only highlights the advancing capabilities of AI but also underscore the need for researchers to critically assess its potential and limitations.

Given these evolving capabilities, it is not surprising that educators’ views on AI in writing assessment are also varied. Many already recognize AI as a tool with the potential to enhance student drafting and revision processes (Dobrin, 2023; Warschauer et al., 2023), potentially leveling the playing field for students with limited access to academic resources. However, alongside these optimistic perspectives, educators also express concerns about the ethical use of AI, its influence on teaching practices, and the potential for over-reliance on technology, which could hinder the development of critical thinking and writing skills (Barrett & Pack, 2023; Bender et al., 2021). These differing perspectives reflect the complex considerations involved in integrating AI into educational assessment.

1.2. Statement of the problem

AI tools, such as ChatGPT, have demonstrated capabilities comparable to humans in educational contexts. However, its adoption in academic environments theoretically relies heavily on university instructors. They play a pivotal role in this evolving landscape. Understanding instructors' views on the advantages and limitations of AI in writing assessment, as well as their trust in such technologies, is crucial for informed implementation. However, there is a paucity of research focusing specifically on university instructors' perspectives regarding AI's role in writing assessment.

Furthermore, while AI has shown promise in grading, discrepancies between human and AI writing have been observed. For example, in the aforementioned study by Herbold et al. (2023), in which human-written and AI-generated essays were compared, it was additionally found that "there are significant linguistic differences between humans and AI-generated content" (p. 8). This could have implications for a grading system utilizing AI, and therefore it underscores the need for a deeper understanding of how AI assessments align with that of humans.

1.3. The aim of the thesis

The aim of this thesis is to investigate university instructors' perspectives on the utilization of AI in assessing writing, as well as to compare the grading outcomes between human and AI raters. This dual approach seeks to provide a comprehensive understanding of the current landscape of AI integration in educational assessment, focusing on both educator attitudes and empirical grading correlations.

Firstly, the thesis examines how university instructors¹ perceive and incorporate AI tools in assessing writing. It explores their trust and confidence in AI-generated assessments, perceived effectiveness, ethical considerations, pedagogical implications, and future outlook. Understanding these perspectives is crucial, as instructors' acceptance and effective use of AI can significantly influence its integration into higher educational practices. Previous research has highlighted the importance of educators' attitudes towards AI in shaping its adoption and effectiveness in academic settings (see Lundgren, 2024).

Secondly, the thesis conducts a correlational analysis of grading outcomes between human and AI² raters, through the results of a grading process of 60 argumentative English as a Second Language (ESL) student essays using the IELTS Writing Task 2 Bands (IELTS-WT2B) as the rubric. This analysis aims to determine the reliability and validity of AI as an assessment tool in language education. Recent studies have shown varying degrees of agreement between human and AI grading, underscoring the need for further investigation into AI's efficacy in educational assessments (e.g., Herbold et al., 2023; Lundgren, 2024).

1.4. Research questions

In light of this thesis' aim, the following research questions are formulated to guide the investigation:

1. To what extent do university instructors¹ use AI tools in assessing writing?
2. What are university instructors' perceptions of effectiveness, confidence, and trust of AI tools in assessing writing?
 - 2.1. How do university instructors perceive the ethical and pedagogical implications of using AI tools in assessing writing?
 - 2.2. How do university instructors foresee the role of AI in the future of assessing writing?
3. To what degree do scores assigned by AI raters² correlate to those assigned by human raters in grading argumentative ESL essays?
 - 3.1. To what extent is there consistency or discrepancy between AI and human scores across the criteria of the IELTS Writing Task 2 Bands?

1.5. Limitations of the thesis

While the findings of this thesis may provide valuable insights, several limitations must be considered to contextualize the results.

¹University-level English language instructors.

²A complete list of AI tools used is available at §3.2.

The sample size and representativeness warrant attention. The survey included 96 university instructors from Turkey and the grading analysis was based on 60 argumentative essays written by adult Colombian ESL learners. Although these figures can offer a substantial data pool, they may not adequately reflect the diversity of university instructors' experiences or the variability in students' writing proficiency.

The possibility of survey response bias must not be ignored. Participation was voluntary, which could have introduced self-selection bias. Instructors with particularly strong views on AI—whether positive or negative—might have been more likely to participate.

Reliance on a small number of human raters could have introduced variability. Only 4 raters with an English as a Foreign Language background participated, which could have potentially limited the generalizability of the findings. Differences in individual grading standards and interpretations of IELTS Writing Task 2 Bands may have influenced the consistency and reliability of human assessments.

Performance of AI tools is influenced by technological advancements and constraints. This thesis' findings should be understood within the context of AI tools' capabilities at the time of research. For instance, the Claude 3.7 Sonnet large language model was released *during* the writing of this thesis.

1.6. List of definitions

In this thesis, several terms are frequently used. For clarity and consistency, they are defined as follows:

1. *Argumentative Essay*: A type of essay that presents a claim or thesis and supports it with logical reasoning and evidence (Hyland, 1990; Weyand et al., 2018). It typically requires the writer to analyze different perspectives and argue for a particular stance.
2. *Artificial Intelligence (AI)*: A branch of computer science focused on creating systems capable of performing tasks that typically require human intelligence, such as learning, reasoning, problem-solving, and language understanding (Russell & Norvig, 2022). In the context of this thesis, AI refers to technologies employed in large language models and assessing writing.

3. *Artificial Intelligence Raters (AI Raters)*: In this thesis, the term *AI raters* refers specifically to large language models. These LLMs are employed in the correlational analysis strand within this thesis.
4. *Automated Essay Scoring (AES)*: A technology-based approach to assessing written essays by using algorithms and artificial intelligence (Page, 1966).
5. *Automated Writing Evaluation (AWE)*: A technology-driven system that provides immediate, formative feedback on various aspects of learner writing (e.g., grammar, coherence, and organization), often combining AES with pedagogical feedback strategies (Attali & Burstein, 2006; Shermis & Burstein, 2003).
6. *English as a Foreign Language (EFL)*: The study of English by individuals whose primary language is not English, particularly in regions where English is not the dominant language (Celce-Murcia et al., 2014).
7. *English as a Second Language (ESL)*: The study of English by individuals whose primary language is not English, typically in contexts where English is the dominant or official language of the society (Celce-Murcia et al., 2014).
8. *English Language Teaching (ELT)*: The practice and theory of teaching English to non-native speakers. ELT encompasses various methodologies and approaches aimed at enhancing English proficiency among learners (Celce-Murcia et al., 2014).
9. *IELTS Writing Task 2 Bands (IELTS-WT2B)*: A standardized rubric used in International English Language Testing System (IELTS) to assess writing tasks (IELTS, 2023). It evaluates the criteria Coherence and Cohesion (CC), Grammatical Range and Accuracy (GRA), Lexical Resource (LR), and Task Response (TR). This rubric serves as the benchmark for grading essays in this thesis.
10. *Large Language Model (LLM)*: A subtype of AI models trained on massive amounts of text data to perform a wide range of natural language understanding and generation tasks. LLMs rely on deep learning architectures like the *transformer* to model linguistic patterns (Devlin et al., 2019; Raffel et al., 2023). In this thesis, LLMs are central to the automated scoring process.

CHAPTER II

LITERATURE REVIEW

In this chapter, a comprehensive yet approachable review of the history and current state of automated essay scoring (AES) and the use of artificial intelligence (AI)—particularly large language models (LLMs)—in writing assessment is presented. The review is split into several key sections: An overview of AES and its advancement from early computational models to contemporary deep learning approaches; a look into the emergence of LLMs and the core details; and lastly, the current adoption status of AI in writing assessment and relevant research.

2.1. A historical perspective into automated essay scoring

The origins of AES can be traced back to the pioneering work of Ellis Batten Page in the mid-1960s, when he introduced Project Essay Grade (PEG). Page’s system quantified surface-level textual features—such as word count, average word length, and syntactic markers—to approximate human judgment (Page, 1966, 1968). Although early computing resources limited the sophistication of AES, PEG laid the groundwork for subsequent advancements.

By the 1980s and 1990s, improvements in computational power enabled the development of more sophisticated AES systems, incorporating statistical and natural language processing (NLP) techniques. Systems such as the Intelligent Essay Assessor (IEA) and ETS e-rater moved beyond simple surface metrics by employing latent semantic analysis (LSA) and Bayesian inference to capture deeper linguistic properties (Landauer et al., 1998; Shermis & Burstein, 2003). The commercial system IntelliMetric further refined AES through a “committee of judges” model, aggregating scores based on hundreds of linguistic features (S. Elliot, 2003).

2.1.1. Early developments and integration of natural language processing techniques

Page’s PEG system was among the first attempts to mechanize essay assessment, demonstrating that quantifiable textual features—such as essay length, vocabulary diversity, and syntactic

complexity—could serve as proxies for human judgment and form the conceptual basis for AES research. However, these early models relied primarily on surface-level analysis. A major advancement came with the introduction of LSA by Landauer et al. (1998), which enabled modeling of semantic relationships within texts and allowed systems to assess coherence and content relevance more effectively. Building on this, the IEA applied LSA-based semantic similarity measures to evaluate essays on the basis of conceptual coherence rather than solely structural features (Perelman, 2014).

Concurrently, ETS developed the e-rater, which expanded the scope of linguistic features considered in scoring by integrating both rule-based and statistical NLP and machine learning (ML) techniques. As detailed by Attali and Burstein (2006), ETS e-rater evaluated grammatical accuracy, discourse structure, lexical sophistication, and organizational patterns, thereby enhancing the reliability of machine-generated scores. These innovations in the 1990s and early 2000s significantly improved agreement with human raters, often reaching levels comparable to inter-rater reliability among expert scorers (J. C. Burstein, 2003; Shermis & Burstein, 2003).

2.1.2. Contemporary automated essay scoring: Powered by artificial intelligence

Recent advancements in NLP and deep learning have transformed AES. Modern systems leverage deep neural networks—particularly transformer-based architectures and LLMs—to capture both surface-level and deep semantic features of writing. These models, such as BERT, XLNet, and ChatGPT have demonstrated scoring correlations with human raters that often surpass inter-rater reliability thresholds (Rodriguez et al., 2019; Taghipour & Ng, 2016). To quote Taghipour and Ng (2016) directly, “for some of the prompts, the agreement between our system and the human annotators is even higher than the agreement between human annotators.” (p. 1889).

Unlike earlier models that relied on handcrafted features, contemporary AES systems automatically extract high-dimensional representations of text, improving generalization across different prompts and writing styles. Additionally, new evaluation techniques have been integrated to enhance grading robustness, ensuring that scores remain stable even when essays include strategically modified content (Kabra et al., 2021).

Hybrid approaches are also emerging, where LLMs assist human raters by flagging inconsistencies and providing detailed feedback. Such collaborations improve grading efficiency and consistency, particularly in descriptive and argumentative writing assessments (see persuasive/expository essays in Xiao et al., 2025, p. 13). Comparative evaluations of AI tools—including various versions of ChatGPT, Google Gemini, Claude AI, and DeepSeek—highlight ongoing improvements in performance and cost-effectiveness (Hardy, 2021).

2.1.3. Challenges, critiques, and future directions

Despite considerable advances in AES systems, persistent concerns remain regarding their validity, reliability, and pedagogical impact. A central critique is that many models rely heavily on surface-level features—such as word counts, lexical sophistication, and syntactic patterns—that may correlate with writing quality but fail to capture deeper dimensions like argument coherence, content relevance, and rhetorical effectiveness. This reliance not only raises doubts about construct validity but also opens avenues for students to *game* the system, artificially inflating scores without substantial improvement in writing skills (Dikli, 2006; Perelman, 2012, 2014; Y. Yang et al., 2002).

Generalizability poses another challenge: scoring models trained on one demographic or instructional context often underperform when applied to different populations, undermining external validity and equity of assessment (Rudner & Liang, 2002; Shermis & Burstein, 2003). Moreover, high correlations frequently reported between AES outputs and human scores may be inflated by the use of homogeneous training and testing datasets drawn from the same source, obscuring true performance in real-world, diverse settings (Wresch, 1993).

Ethical and pedagogical debates further complicate the deployment of AES. Critics argue that automated scoring risks diminishing teachers' roles in offering detailed, personalized feedback—a cornerstone of student writing development—and may perpetuate algorithmic biases if models are trained on unrepresentative corpora, inadvertently disadvantaging certain learner groups (Ben-Simon & Bennett, 2007; Warschauer & Ware, 2006; Warschauer et al.,

2023). In high-stakes contexts, fairness concerns and the potential for score inflation have spurred calls for more transparent, accountable systems (Perelman, 2014).

Transparency and explainability remain critical issues. Traditional AES approaches, often based on linear regression or LSA, function as “black boxes,” offering little insight into decision rationales. While recent deep learning methods promise richer semantic and contextual modeling, their complex architectures can further obscure interpretability, raising accountability challenges and hesitancy among educators (Ke & Ng, 2019; Kim & Jo, 2024; Taghipour & Ng, 2016).

2.1.4. Automated writing evaluation and its relationship to automated essay scoring

While AES has primarily focused on generating scores that approximate human judgment, automated writing evaluation (AWE) emerged as a broader pedagogical framework encompassing not only scoring but also comprehensive feedback provision for writing instruction (Stevenson, 2016; Warschauer, 2010). AWE systems integrate various computational techniques to analyze student writing and provide actionable feedback on multiple dimensions, including content development, organization, style, mechanics, and language conventions (Cotos, 2014; Ranalli, 2018).

The development of AWE can be traced to the recognition that scoring alone provides limited instructional value for students seeking to improve their writing abilities. Early AWE systems such as *Writer’s Workbench*, developed at Bell Laboratories in the 1980s, pioneered the concept of providing detailed diagnostic feedback on stylistic and mechanical aspects of writing (Macdonald et al., 1982). This pedagogical orientation distinguished AWE from pure AES applications, emphasizing formative assessment over summative evaluation (Grimes & Warschauer, 2010; Roscoe & McNamara, 2013).

Contemporary AWE platforms—including Criterion, WriteToLearn, and Revision Assistant—combine sophisticated AES engines with feedback generation modules that target specific writing weaknesses (J. Burstein, 2012; Foltz & Rosenstein, 2015; J. Wilson & Czik, 2016). These systems typically employ NLP techniques to identify patterns in student writing and generate targeted suggestions for improvement. For instance, Criterion leverages ETS

e-rater's scoring capabilities while additionally providing feedback on grammar, usage, mechanics, style, and organization through its diagnostic modules (Attali, 2004; Chodorow et al., 2010).

The relationship between AWE and AES is both complementary and complex. While AES forms the analytical core of many AWE systems, AWE extends beyond scoring to address pedagogical concerns about writing development (Shermis & Wilson, 2024; Stevenson, 2016). This distinction has important implications for implementation and evaluation. AWE systems must balance accuracy in assessment with the pedagogical utility of feedback, often requiring trade-offs between scoring precision and instructional clarity (Link et al., 2022; Ranalli, 2018).

Research has demonstrated that AWE can positively impact student writing performance when integrated thoughtfully into instructional contexts. Meta-analyses by Stevenson and Phakiti (2019) and J. Wilson and Roscoe (2020) found moderate effect sizes for AWE interventions on writing quality, particularly when combined with teacher instruction and peer feedback. However, the effectiveness of AWE depends heavily on implementation factors, including student training, teacher support, and alignment with curricular goals (Kellogg et al., 2010; Warschauer & Grimes, 2008).

Despite these advances, AWE faces several criticisms that echo concerns about AES while introducing new pedagogical challenges. Critics argue that AWE feedback often remains superficial, focusing disproportionately on surface-level errors rather than higher-order concerns such as argumentation and audience awareness (Deane, 2013; Vojak et al., 2011). Additionally, the generic nature of automated feedback may fail to address individual student needs or disciplinary writing conventions (Cotos, 2011; Ranalli, 2021). Studies have also raised concerns about student over-reliance on AWE systems, potentially hindering the development of self-assessment skills and metacognitive awareness essential for writing development (Li et al., 2017; Saricaoglu & Bilki, 2021).

Furthermore, the integration of AWE into writing pedagogy raises questions about the changing role of writing instructors. While proponents argue that AWE can free teachers to focus on higher-order instruction, critics worry about the potential displacement of human judgment and the detailed feedback that experienced instructors provide (Chapelle, 2017; Warschauer,

2010). This tension reflects broader debates about technology's role in education and the balance between efficiency and pedagogical effectiveness (N. Elliot, 2016; Ericsson & Haswell, 2006).

2.2. The beginning of large language models

The development of LLMs marks a significant turning point in AI and NLP. Early computational efforts to model human language relied on probabilistic approaches such as *n-grams*, which predicted word sequences based on statistical likelihood (Deller & Hansen, 2005). However, these models were constrained by limited computational power, sparse datasets, and a lack of ability to account for contextual information beyond fixed windows of words (Jurafsky & Martin, 2024).

A breakthrough came with the advent of neural networks in the late 1990s and early 2000s, enabling the creation of distributed word representations such as word embeddings. These embeddings, exemplified by Mikolov et al.'s (2013) *Word2Vec*, captured semantic relationships between words. The introduction of recurrent neural networks (RNNs) and their variants like long short-term memory (LSTM) provided mechanisms to maintain contextual information over longer spans. Despite these advancements, RNNs were hampered by vanishing gradient issues and inefficiencies in capturing relationships across long distances in text.

The emergence of the *transformer* architecture, introduced by Vaswani et al. (2023) in their seminal work "Attention Is All You Need", marked the true advent of modern LLMs. The self-attention mechanism within transformers allowed models to weigh the importance of all words in a sentence relative to each other, addressing the challenges posed by RNNs and LSTMs. This advancement enabled the modeling of long-range dependencies and parallelized computation, vastly improving training efficiency.

Building on this foundation, researchers recognized that increasing the number of parameters and the size of training datasets led to significant performance improvements. This scaling paradigm facilitated the development of models capable of generating human-like text with remarkable fluency and contextual understanding. Notable milestones in this trajectory include OpenAI's generative pre-trained transformer (GPT) series and Google's bidirectional encoder

representations from transformers (BERT), both of which demonstrated the transformative potential of pre-trained models fine-tuned for specific tasks (Devlin et al., 2019; Radford et al., 2018). These developments underscored the importance of pre-training on large, diverse text corpora to enhance downstream task adaptability.

2.2.1. The transformer deep learning architecture

Expanding upon the discussion of the transformer model, its introduction by Vaswani et al. (2023) represented a pivotal advancement in NLP by replacing traditional recurrent structures with an attention-based mechanism. Unlike RNNs and LSTMs, transformers process input sequences in parallel, enabling more efficient capture of long-range dependencies.

At its core, the transformer relies on a multi-head self-attention mechanism, which dynamically assigns importance to different words within a sequence. This approach allows the model to learn distinct, complementary representations across different subspaces. Additionally, position-wise feedforward neural networks, residual connections, and layer normalization contribute to stable gradient flow during training. Since transformers do not inherently encode sequential order, positional encodings are incorporated to provide necessary order information (Vaswani et al., 2023).

The architecture consists of an encoder and a decoder. The encoder transforms input token sequences into contextualized representations, while the decoder generates output sequences using masked self-attention to maintain auto-regressive properties. Notably, BERT leverages the transformer encoder to generate deeply bidirectional representations, demonstrating its effectiveness in various NLP tasks (Devlin et al., 2019).

2.2.2. Generative pre-trained transformers

Building upon the transformer model, OpenAI's GPT series exemplifies the potential of transformer-based architectures in NLP. Introduced with GPT-1 (Radford et al., 2018), this series established a paradigm of pre-training on extensive text datasets followed by task-specific fine-tuning.

Pre-training involves unsupervised learning from large corpora, with objectives such as causal language modeling. Fine-tuning then adapts the model for specific tasks. This framework, refined in GPT-2 and GPT-3, demonstrated significant improvements in coherence and contextual awareness (Brown et al., 2020). GPT-3's ability to perform few-shot and zero-shot learning further highlighted the advantages of large-scale pre-training.

Despite their capabilities, GPT models have raised ethical concerns, including issues of bias, misinformation, and environmental impact. Addressing these challenges necessitates fairness-aware training processes and efficient model architectures (Bender et al., 2021).

2.2.3. Zero-shot in large language models

Zero-shot learning in LLMs extends the pre-training paradigm by allowing models to perform tasks without explicit task-specific training. This capability arises from the extensive knowledge acquired during pre-training (Brown et al., 2020). Instead of relying on annotated datasets, zero-shot models generate responses based on natural language instructions that encapsulate task requirements.

Research has demonstrated that careful prompt design can enhance zero-shot performance. For example, Kojima et al. (2023) found that adding trigger phrases such as “Let’s think step by step” (p. 6) improved complex reasoning tasks. This technique, known as *chain-of-thought prompting* (Wei et al., 2023, pp. 2–3), highlights the latent reasoning capabilities within LLMs.

In education, zero-shot learning offers potential applications in automated assessment. By aligning prompts with established grading rubrics, educators can leverage LLMs for scalable and cost-effective evaluation. This flexibility makes zero-shot models particularly suitable for writing assessment, where a unique judgment is required. Continued research in prompt engineering and model architectures could refine zero-shot capabilities, broadening applicability in educational settings.

2.2.4. Large language models’ emergence in education

The integration of LLMs into education marks a significant evolution in assessment methodologies. Building upon prior discussions of their linguistic and computational

capabilities, LLMs have proven particularly effective in AWE.

Empirical studies have demonstrated that LLMs align with established grading rubrics, providing consistent and rapid feedback (Holmes et al., 2019). The efficiency of AI-based grading, compared to traditional human assessment, has led to increased adoption in higher education settings (Zawacki-Richter et al., 2019). However, discrepancies between AIs and human raters necessitate systematic evaluations to ensure fairness and reliability.

Beyond grading, the rise of LLMs in education is part of a broader digital transformation that emphasizes personalized instruction and data-driven decision-making. AI agents are increasingly seen as complementary tools that enhance, rather than replace, human judgment. This shift has prompted investigations into the ethical and pedagogical implications of AI in education. Key concerns include bias, transparency, and accountability (Pedro et al., 2019).

2.3. Use of artificial intelligence in assessment: English writing context

The incorporation of AI in writing assessment has gained traction in the recent decades. This section reviews current approaches, emphasizing three interconnected aspects: methodological frameworks, comparative grading between humans and AI raters, and educators' perceptions regarding AI deployment in assessment contexts.

2.3.1. Methodological frameworks in writing assessment

Historically, writing assessment frameworks have aimed to capture the multi-faceted nature of writing proficiency while maintaining clarity, consistency, and accuracy in evaluating student outputs (Hyland, 2019). Within English as a Foreign Language (EFL) settings, methodological frameworks typically emphasize criteria that assess rhetorical structure, linguistic accuracy, coherence, and content relevance in students' compositions (Bazerman et al., 2012; Hyland, 1990). In particular, rubric-based approaches have gained prominence, as they anchor evaluators' scoring decisions to explicit, structured criteria with clear justifications (Andrade, 2000; Diab & Balaa, 2011).

Rubrics primarily offer criterion-referenced guidelines for two purposes: providing instructional guidance to support student learning, and ensuring standardized, reliable evaluation across assessors (Diab & Balaa, 2011). Clearly articulated expectations enable learners to align their writing process with task requirements and consequently enhance their skills through iterative drafting and revision cycles (Andrade, 2000). Simultaneously, educators benefit from structured objectivity, facilitating robust inter-rater reliability and comparative insights across various learner populations (Shabani & Panahi, 2020). Despite these advantages, critics argue that rubric-based approaches risk oversimplifying the complexity and contextual subtlety inherent in writing (Broad, 2003; M. Wilson, 2006, 2018).

Within standardized academic assessments, such as IELTS Writing Task 2 Bands (IELTS-WT2B), explicit rubrics guide scoring practices by clearly defining core evaluative parameters. In the case of IELTS-WT2B, these are Coherence and Cohesion, Grammatical Range and Accuracy, Lexical Resource, and Task Response (Cotton & Wilson, 2011; IELTS, 2023, 2024b). Traditionally, evaluation according to these rubrics has heavily relied upon trained human raters whose calibration and training routines ensure consistency across scoring (Cotton & Wilson, 2011; IELTS, n.d.). These standardized frameworks have thus shaped global expectations regarding English writing proficiency within international university admissions contexts (British Council, 2024; IELTS, 2024a, 2024c).

Recently, the rapid advancement of AI technologies has facilitated their integration into rubric-based evaluation models, introducing novel methodological paradigms in writing assessment. Automated systems such as IntelliMetric, ETS e-rater, and IEA employ linguistic computational techniques designed to replicate and complement human evaluation capabilities within rubric-informed structures (Attali & Burstein, 2006; Dikli, 2006; S. Elliot, 2003; Foltz et al., 2013; Hussein et al., 2019). By leveraging NLP, AI tools can extract evaluative patterns from extensive corpora of human-scored texts, facilitating consistent analysis of student work against complex linguistic criteria (J. C. Burstein, 2003; Kabra et al., 2021; Ke & Ng, 2019).

Despite these technological advancements, significant challenges remain in adapting AI tools to assess comprehensive, qualitative aspects of academic writing, including critical thinking, originality, creativity, and higher-order argumentation skills, which traditionally resist straightforward quantification (Hyland, 1990; Perelman, 2012, 2014; Warschauer &

Ware, 2006; Weyand et al., 2018). Consequently, scholarly and professional consensus highlights the necessity of complementary assessment practices integrating both AI and human raters, thereby ensuring robust validity and continuous methodological enhancements across multidimensional writing constructs (Warschauer & Ware, 2006; Williamson et al., 2012; Y. Yang et al., 2002).

2.3.2. Human and artificial intelligence grading

Advancements in computational linguistics have empowered AES systems as effective supplements to traditional human assessments, particularly in large-scale contexts demanding efficiency and scalability (J. C. Burstein, 2003; Dikli, 2006; Godwin-Jones, 2022; Ke & Ng, 2019). Prominent automated tools—including ETS e-rater, Pearson’s IEA, and Vantage’s IntelliMetric—illustrate how computational models attempt to emulate human grading patterns by analyzing linguistic, textual coherence, and rhetorical features extracted from human-evaluated benchmark essays (Attali et al., 2010; Foltz et al., 2013; Shermis & Burstein, 2003; Williamson et al., 2006).

Human raters contribute critical cognitive judgments anchored in expertise to identify content depth, originality, thematic consistency, logical coherence, and contextual appropriateness (Hyland, 2019; Weyand et al., 2018). Yet human evaluation approaches face persistent challenges related to subjectivity, reliability limitations, rater fatigue, and resource intensiveness (Hussein et al., 2019; Shermis & Burstein, 2003). Additionally, maintaining consistency across human raters involves substantial ongoing training procedures and calibration protocols, adding considerable logistical complexity and cost (IELTS, 2024b; Shabani & Panahi, 2020).

Conversely, AI-based grading directly addresses these limitations by offering enhanced efficiency, scalability, objectivity, and consistent evaluation standards across a large throughput of student compositions (Gardner et al., 2021; Warschauer & Ware, 2006). Research demonstrates strong agreements between AI and human scores for surface-level textual features such as grammar, syntax, spelling, and holistic coherence (Attali & Burstein, 2006; Landauer et al., 2002; Ramesh & Sanampudi, 2022; Shermis, 2014). However, evaluating

more complex cognitive skills—such as detailed argumentation, effective critical thinking, cohesive originality, and higher-order reasoning—remains difficult for automated systems and continues to predominantly rely on human raters’ insights (Bouziane & Bouziane, 2024; Kabra et al., 2021; Perelman, 2012, 2014; K. Yang et al., 2024). Furthermore, algorithmic biases inherent in AI scoring may disadvantage particular sociodemographic groups due to biases embedded in training corpora used for development (Baker & Hawn, 2022; Bender et al., 2021; K. Yang et al., 2024).

Empirical research recommends a hybrid approach drawing upon strengths inherent in both human and AI evaluation methods (Kemal & Liman-Kaban, 2024; Shermis & Hamner, 2013; Williamson et al., 2012). Studies consistently indicate that blended approaches not only enhance assessment accuracy in encompassing multidimensional writing skills, but also improve scoring reliability and generalizability across evaluative contexts (Seßler et al., 2024; Wetzler et al., 2024; Zhong et al., 2024). Reflecting these advantages, standardized assessments such as International English Language Testing System (IELTS), Test of English as a Foreign Language (TOEFL), and Graduate Record Examinations (GRE) have increasingly adopted hybridized evaluation frameworks (Attali et al., 2010; IELTS, 2024b).

2.3.3. Educators’ perceptions of artificial intelligence in assessment

Educators hold diverse perceptions regarding AI’s assessment potential, shaped significantly by its efficiency, academic integrity implications, impact on student learning, and relevance for pedagogical practice. Attitudes toward AI assessment tools range from strong endorsement of their efficiency to expressions of concern regarding their effects on students’ critical thinking abilities and originality (Alharbi, 2023; Jayaron & Blessy, 2024).

A central concern among educators is the potential threat to academic integrity posed by AI tools in writing tasks. These tools have the potential to detect plagiarism and provide targeted language revisions; however, educators caution that over-reliance may undermine deeper engagement and originality in composing written works (Leah et al., 2024). Moreover, AI’s predominantly surface-level feedback could potentially diminish student motivation to

develop sophisticated arguments or critically reflective positions (Lukianenko & Kornieva, 2024).

Nevertheless, educators widely acknowledge AI's ability to enhance grading efficiency and maintain evaluation consistency, significantly reducing grading workloads and minimizing biases common in solely human assessments (Nisar Ahmad Koka, 2023). Yet challenges arise regarding AI's capability to evaluate higher-order skills, such as sophisticated argumentation or creativity, which remains better suited to human judgment (Kizilcec et al., 2024).

Reliability comparisons between AI and human raters remain contentious. While AI increasingly demonstrates alignment with human evaluative patterns, critics argue that it lacks the contextual distinction and pedagogical discernment inherent to human raters (Merkviladze, 2024; Mulyono, 2023). Furthermore, Roe et al. (2024) identified notable differences in acceptance of AI tools among educators and students, reflecting broader perception gaps.

CHAPTER III

METHODOLOGY

3.1. Research design

This thesis adopts a concurrent multimethod quantitative research design, in which two (mostly-)quantitative strands are implemented simultaneously and their findings subsequently triangulated to address the central research questions (as defined by Brewer & Hunter, 2006). Multimethod research—distinct from mixed methods designs—entails the use of more than one quantitative or qualitative method, without requiring both paradigms in a single construct (as explained by Johnson & Christensen, 2014). Within this thesis, the first research strand captures university instructors' perceptions via a descriptive cross-sectional survey design, while the second research strand examines empirical score variability through a correlational research.

The first research strand employs a descriptive, cross-sectional survey approach. A survey research involves quantitative measurements of “trends, attitudes, or opinions” held by a larger group through examination of a representative sample of that population (Fowler, 2008, as cited in Creswell and Creswell, 2018, pp. 49–50), while cross-sectional designs specifically capture these data simultaneously rather than over intervals (Levin, 2006).

Concurrently, the second research strand utilizes a correlational research approach. Correlational research quantifies the degree and direction of association between two or more variables without experimental manipulation, thereby identifying patterns or relationships within naturally occurring data (Creswell, 2013, p. 338).

Results from each strand were analyzed independently and then merged to explore alignments and divergences between perceived and empirical fairness in AI-assisted assessment. By integrating a descriptive survey component with a correlational analysis, this research achieves *holistic* triangulation—“the use of multiple measures [to] uncover some unique variance which otherwise [would] have been neglected by single methods” (Jick, 1979, p. 603)—thereby enhancing the credibility, validity, and comprehensiveness of its findings.

3.2. Participants

Three distinct yet complementary cohorts contributed data to the thesis: (a) university instructors who assess English writing, (b) adult Colombian learners of English—explained further in §3.3.—and (c) four English as a Foreign Language (EFL) academicians, who graded the essays of the Colombian learners.

Within the university instructors cohort, participants were recruited via purposive sampling to secure a pool of instructors with direct, recent experience in writing assessment. Purposive sampling is a non-probability sampling technique in which cases are deliberately selected because they possess the characteristics required to address the research questions (Etikan, 2016; Palinkas et al., 2015). Eligibility required at least one prior instance of grading student essays in English.

The resulting group comprised 96 participants drawn primarily from English Language Teaching (ELT) and Applied Linguistics (37.50% and 21.87%, respectively) programs at different universities in Turkey. The participants represented a wide span of academic ranks (from research assistants to full professors) and career stages (1 to 15+ years of experience), with the modal age band falling between 31 and 40 years. Complete numerical information are presented in Figures 3.1 and 3.2.

The second cohort consisted of 60 argumentative English as a Second Language (ESL) essays written by Colombian adult learners of English and archived in “The Small Corpus of Colombian English as a Second Language Essays” (published by Velasco, 2023). Each essay was scored by two distinct evaluator groups: (1) four humans and (2) twenty-five artificial intelligence (AI) tools.

The human raters are full-time university instructors working within ELT programs. The instructors were selected because they were accessible by the researcher, akin to convenience sampling. Convenience sampling is a type of non-probability sampling in which participants are chosen based on their ease of access, typically selected from a population that is readily available to the researcher (Etikan, 2016). The human raters’ background are given below:

Rater 1. Turkish; Bachelor of Arts (B.A.) in ELT; taught grammar and the four language skills in undergraduate programs since 2012.

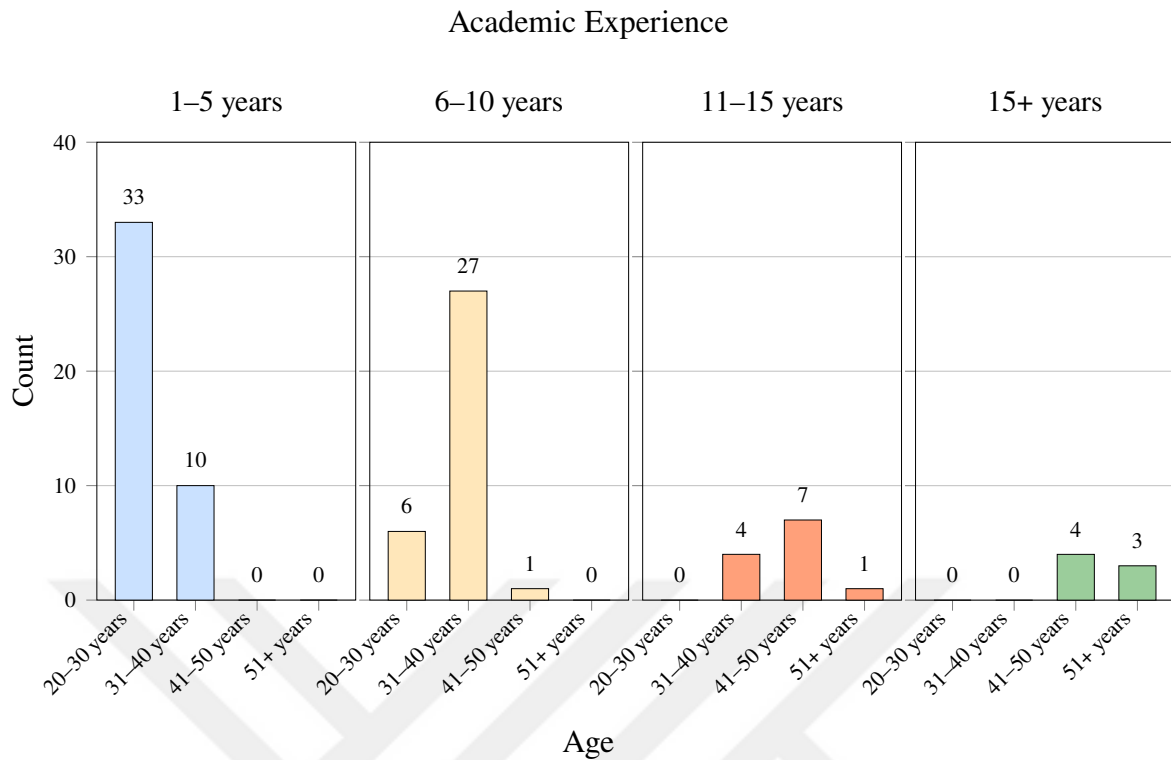


Figure 3.1

Distribution of Participants' Academic Experience by Age Group

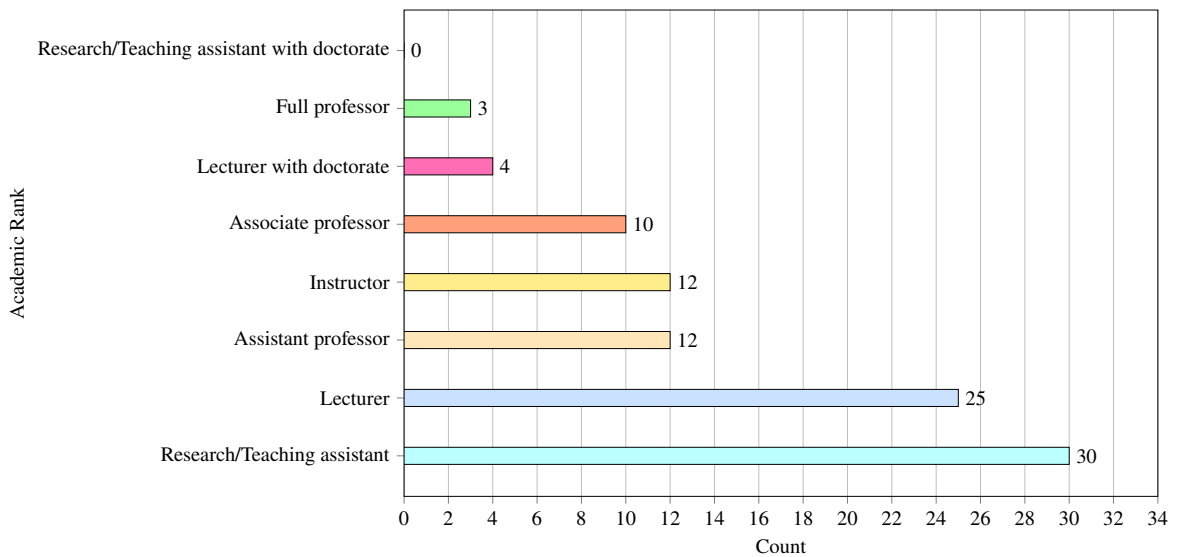


Figure 3.2

Distribution of Participants' Academic Ranks

- Rater 2. Turkish; B.A. in ELT; Master of Arts (M.A.) in Translation Studies; teaching English writing and preparatory courses since 2014.
- Rater 3. Turkish; B.A. in ELT; M.A. in Educational Sciences; doctoral candidate in Educational Sciences; teaching preparatory English since 2023.
- Rater 4. Turkish; B.A. in ELT; M.A. in ELT; doctoral candidate in ELT; teaching practicum courses since 2023.

The AI tools were chosen based on the public availability of information regarding their capabilities and reputations (list available in Table 3.1). Tools developed by well-established organizations such as Anthropic, DeepSeek, OpenAI, and Google were prioritized, as these are potentially already widely recognized within academic and professional ELT contexts, making them familiar to most university instructors. Although the inclusion of other local or open-weight AI tools—such as Llama, Mistral, and Qwen—was initially considered, the researcher was unable to implement them due to limited access to the required computational resources and financial constraints.

Table 3.1
List of AI Tools Used as AI Raters

Descriptor Given to APIs ^a	Model Name
claude-3-5-haiku-20241022	Anthropic Claude 3.5 Haiku
claude-3-5-sonnet-20240620	Anthropic Claude 3.5 Sonnet (Old)
claude-3-5-sonnet-20241022	Anthropic Claude 3.5 Sonnet (New)
claude-3-7-sonnet-20250219-thinking	Anthropic Claude 3.7 Sonnet (Thinking Enabled)
claude-3-7-sonnet-20250219	Anthropic Claude 3.7 Sonnet (Thinking Disabled)
claude-3-haiku-20240307	Anthropic Claude 3 Haiku

Continued on next page

Table 3.1*List of AI Tools Used as AI Raters (Continued)*

Descriptor Given to APIs ^a	Model Name
claude-3-opus-20240229	Anthropic Claude 3 Opus
deepseek-chat	DeepSeek-V3-0324
deepseek-reasoner	DeepSeek-R1 ^b
gemini-1.5-flash-8b	Google Gemini 1.5 Flash-8B
gemini-1.5-flash	Google Gemini 1.5 Flash
gemini-1.5-pro	Google Gemini 1.5 Pro
gemini-2.0-flash-lite	Google Gemini 2.0 Flash-Lite
gemini-2.0-flash	Google Gemini 2.0 Flash
gemini-2.5-flash-preview-04-17	Google Gemini 2.5 Flash Preview 04-17
gemini-2.5-pro-exp-03-25	Google Gemini 2.5 Pro Experimental 03-25
gpt-4.1-2025-04-14	OpenAI GPT-4.1
gpt-4.1-mini-2025-04-14	OpenAI GPT-4.1 mini
gpt-4.1-nano-2025-04-14	OpenAI GPT-4.1 nano
gpt-4o-2024-11-20	OpenAI GPT-4o
gpt-4o-mini-2024-07-18	OpenAI GPT-4o mini
o1-2024-12-17	OpenAI o1 ^c
o3-2025-04-16	OpenAI o3 ^c
o3-mini-2025-01-31	OpenAI o3-mini ^c
o4-mini-2025-04-16	OpenAI o4-mini ^c

^a As documented by the respective application programming interfaces.^b Labelled “DeepSeek-R1 Release 2025/01/20” in the website documentation.^c Reasoning effort was set to *medium*.

3.3. Data collection instruments

To address the research questions, two complementary instruments were employed: (a) a structured survey that elicited university instructors' perceptions of AI tools in writing assessment and (b) a rating task in which both humans and AI raters graded argumentative ESL essays drawn from a publicly available corpus. The following paragraphs describe each instrument, the constructs it operationalizes, and the rationale for their selection.

The survey captured five theoretically grounded constructs: AI trustworthiness, perceived effectiveness, pedagogical alignment, ethical acceptability, and future outlook. Each construct was measured with multiple five-point Likert-type items. A Likert scale is a symmetrical, equidistant series of response options—typically ranging from strong disagreement to strong agreement—designed to quantify respondents' attitudes or beliefs (“the simpler technique” described in Likert, 1932). Five-point formats maximize reliability while minimizing respondent fatigue and central-tendency bias (Joshi et al., 2015). Because only extreme options were verbally labelled, the unlabeled midpoint functioned as an *implied neutral* category, an approach shown to reduce acquiescence bias without compromising scale validity (Chyung et al., 2017; Kulas et al., 2008; Nadler et al., 2015). In addition to close-ended items, optional open-ended items invited narrative comments, thereby enriching quantitative findings with qualitative data.

The survey instrument underwent rigorous development and validation procedures. During the development phase, expert opinions were solicited from two distinguished academics: one specializing in assessment and evaluation, and another in ELT. Their critical feedback informed revisions to item phrasing, construct coverage, and response scale clarity. Mainly, Likert-scale item definitions were adjusted, possible bias in all items were rechecked, and open-ended items were added to mitigate potential risks of data loss. Following these modifications, a pilot application was conducted with four university instructors actively engaged in evaluating student English writing. This iterative process identified necessary refinements to question comprehensibility and procedural instructions, which were implemented prior to full deployment. Finally, ethical approval for administering the instrument was formally granted by the Sakarya University Ethics Committee, confirming adherence to established protocols for research involving human participants.

The second instrument involved double-rating a subset of 60 argumentative ESL essays that were randomly sampled from “The Small Corpus of Colombian English as a Second Language Essays (SCoCESLE)” (published by Velasco, 2023). This corpus was selected for several methodological reasons that aligned with the research objectives. First, it directly addresses the research questions concerning the grading of argumentative essays written in English by non-native speakers. Second, the corpus is readily and easily accessible online at no cost, facilitating research reproducibility. Third, it provides a substantial dataset of 272 argumentative essays that have already been anonymized, ensuring both statistical robustness and ethical compliance. Fourth, the essay prompts are designed in a similar manner to International English Language Testing System (IELTS) Writing tasks, with the corpus documentation explicitly noting that topics were “randomly chosen from a list of 31 IELTS-like essays” (Velasco, 2023, p. 7). Finally, the corpus documentation includes Common European Framework of Reference for Languages (CEFR) proficiency descriptions for each essay, providing essential baseline data for analysis. According to the corpus documentation, writers’ proficiency levels range from A2 to C1 on the CEFR scale (Council of Europe, 2020), an internationally recognized continuum of communicative competence from basic (A1) to proficient user (C2).

Essays were scored with the IELTS Writing Task 2 Bands (IELTS-WT2B), a criterion-based rubric widely adopted in high-stakes language testing and second language writing research because of its demonstrated reliability, construct validity, and international recognition (Hawkey, 2006; Shabani & Panahi, 2020). The rubric operationalizes four constructs (defined in IELTS, 2024b, 2024d), each rated on a 0 to 9 scale:

1. *Coherence and Cohesion (CC)*: Evaluation of the overall organization and logical progression of ideas. Here, *coherence* concerns the logical sequencing of information and arguments, and *cohesion* concerns the varied and appropriate deployment of linking devices (e.g. conjunctions, pronouns, logical connectors) to clarify relationships between and within sentences.
2. *Grammatical Range and Accuracy (GRA)*: Assessment of the variety and correct application of grammatical structures at the sentence level.
3. *Lexical Resource (LR)*: Assessment of the breadth of vocabulary and the precision and

suitability of word choice relative to the task requirements.

4. *Task Response (TR)*: Degree to which the writing presents and sustains a clear position in response to the prompt, in at least 250 words, with arguments substantiated by evidence or relevant personal examples.

Using the IELTS-WT2B allows the present research to (a) benchmark human and *machine* ratings against a well-established professional standard, (b) facilitate comparison with prior studies that employ the same metric, and (c) offer practitioners easily interpretable results that map onto familiar score bands.

The four human raters applied the rubric after a brief norming session, and each of the AI tools was prompted with identical rubric wording. The resulting dataset contains analytic scores for all four criteria plus a composite band (the sum of all criteria), enabling fine-grained analysis of inter-rater agreement across human and AI modalities.

These methods align with the recent literature highlighting the emerging role and challenges of AI in assessment tasks (e.g., Flodén, 2024; Lundgren, 2024; Seßler et al., 2024) and seek to contribute empirical evidence on the practical viability of integrating AI tools into assessment tasks.

3.4. Data collection procedure

Data were gathered concurrently between January and May 2025. All data collection tasks were planned, piloted, and executed under a single ethical protocol approved by the Sakarya University Ethics Committee.

The survey phase began with institutional negotiations. In January 2025, the researcher contacted the institutional correspondence offices of 30 randomly selected universities in Turkey to obtain written permission to approach their academic staff. Once faculty-level approvals were secured, each participating institution shared the Google Forms link to its English language instructors. The survey link remained active until the end of May 2025, producing a time window long enough to accommodate differing academic calendars. All responses were automatically timestamped and stored locally; no identifying information was collected.

During the survey data collection, 60 argumentative ESL essays (retrieved from the corpus SCoCESLE) were uploaded to separate Google Classroom instances. The four ELT instructors were then invited to serve as voluntary raters. During a brief orientation, raters received (a) an aim overview, (b) the IELTS-WT2B, and (c) instructions on how to record scores. No information about the writers or the corpus was disclosed, thereby reducing expectancy effects. Raters had thirty days to evaluate all essays independently, after which raw scoresheets were exported for analysis.

For the ratings by AI tools, 25 large language models (LLMs)—comprising both commercial and open-source models—were queried with an identical prompt (available in §A) that embedded the IELTS-WT2B verbatim in plain text. The prompt was intentionally concise so as to reduce extraneous variability and to approximate the minimal instructions typically given by an average user (Knoth et al., 2024; Zamfirescu-Pereira et al., 2023). Structured JSON output was enforced whenever the application programming interfaces (APIs) allowed; for models lacking native schema support (e.g., DeepSeek, Claude 3.7 Sonnet), the rubric was supplemented with an explicit JSON template or tool definition to ensure output consistency. Core generation parameters (*temperature* and *top-p* set to 1; *frequency* and *presence penalty* set to 0) were held constant across all models, thereby isolating model configuration as the primary source of variation. Batch processing was used when supported by the provider; otherwise, essays were prompted sequentially. All AI rater runs were finished by the first week of May 2025.

While the prompt design adhered to a minimal prompt-engineering paradigm, it is important to acknowledge that prompt structure itself can materially influence AI rater behavior. For instance, Stahl et al. (2024) found that leveraging richer instruction templates and integrated feedback prompts led to measurable gains in both scoring consistency and the pedagogical value of generated feedback. Although exploring the full spectrum of prompt-engineering strategies—for example, few-shot injection, *chain-of-thought* scaffolds, or dynamic prompt adaptation—lies beyond the present research’s aims.

Upon completion of both human and AI grading, all datasets were merged into a single file that stored, for each essay, 4 human scores, 25 AI scores, and associated metadata (essay and rater IDs). No personal data were collected at any stage.

3.5. Data preparation and cleaning for analysis

Consistent and transparent data preparation procedures were implemented in order to maximize analytic validity and reproducibility. All steps were documented in a version-controlled repository and executed with open-source Python 3.13 libraries. Prior to any transformation, raw files were retained in read-only archives, thereby preserving an immutable record of the original observations.

For the survey of university instructors, the raw export from Google Forms was screened for duplicate submissions and straight-lining patterns; none were detected. The dataset contained complete responses from all participating university instructors with no missing data identified during the initial data examination. To facilitate statistical analysis, all categorical response items were systematically recoded into integer values. This transformation process maintained the relationships inherent in the original response scales while enabling the application of appropriate analytical techniques. The coding scheme was consistently applied across all survey items to ensure uniformity in the dataset structure. Data validation procedures confirmed the integrity of the recoded variables, with no inconsistencies or anomalies detected in the final dataset. The cleaned dataset was organized in a format suitable for statistical analysis.

For human and AI scores, raw scoring files—initially stored in mixed CSV and JSONL formats—were concatenated, indexed, and converted into both wide (essays as rows, scores as columns) and long (essay-rater pairs as rows, scores as columns) tables using a custom Python script. Data-type validation checks ensured that every criterion score fell within the IELTS-WT2B range (0 to 9). No missing values were identified. To facilitate correlational analyzes, a binary flag (*is_human_grader*) distinguished human raters (value = 1, $N = 4$) from AI raters (value = 0, $N = 25$). Variables including the arithmetic means of each criterion were then appended for each essay, *HUMAN_MEAN* for the human scores and *AI_MEAN* for the AI scores.

Both cleaned datasets were saved in UTF-8 encoded CSV files, accompanied by detailed data dictionaries specifying variable labels, coding schemes, and permissible value ranges.

3.6. Data analysis procedures

All statistical analyzes were carried out in Python 3.13. Data wrangling and descriptive statistics relied on *pandas* and *numpy*; reliability coefficients on *pingouin* and *krippendorff*; t-tests, Wilcoxon signed-rank tests, correlation coefficients and Shapiro-Wilk tests on *scipy* and *statsmodels*; and visualisation on *seaborn* and *matplotlib*. Analytic notebooks are archived in the project repository to ensure full reproducibility.

The first research strand combined close-ended items and open-ended comments. Internal consistency of the multi-item attitude scales was evaluated with Cronbach's alpha (C_α). C_α represents the mean of all split-half reliabilities, thus estimating the proportion of true score variance captured by the scale (Cronbach, 1951). Values above 0.70 are generally considered adequately reliable for research purposes, with values above 0.80 signaling good, and values above 0.90 excellent reliability (Tavakol & Dennick, 2011). Qualitative comments were subjected to inductive content analysis: codes were generated iteratively, clustered into categories, and refined until theoretical saturation was reached (Morin et al., 2021; Saldaña, 2021).

The second research strand consisted exclusively of quantitative ratings assigned to 60 argumentative ESL essays by 4 human raters and 25 AI raters. Reliability for each IELTS-WT2B criterion and for the total score was quantified with two complementary indices: Krippendorff's alpha (K_α) and intraclass correlation coefficient (ICC_{2k}). K_α is a versatile coefficient applicable to any number of raters and various levels of measurement; it captures the degree to which observed disagreement exceeds that expected by chance (Hayes & Krippendorff, 2007; Krippendorff, 2019). Bootstrap resampling (1000 iterations, 95% CI) provided precision estimates for K_α . Since more than one rater scored the same essay, ICC_{2k} was utilized. ICC_{2k} expresses the ratio of object variance to total observed variance and is appropriate when raters are sampled from a broader population and mean scores across k raters are analyzed (Shrout & Fleiss, 1979). Benchmarks adopted here classify ICC_{2k} values below 0.50 as poor, values between 0.50 and 0.75 as moderate, values between 0.75 and 0.90 as good, and values above 0.90 as excellent (as defined by Koo & Li, 2016).

To compare arithmetic means between human and AI scores, normality of the paired difference

scores was first verified using the Shapiro-Wilk test. The Shapiro-Wilk test evaluates the null hypothesis that a sample is drawn from a normally distributed population by computing the W statistic. Small values of W indicate departures from normality, and significance testing of W determines whether to reject normality at a chosen alpha level (Razali & Yap, 2011; Shapiro & Wilk, 1965; Shapiro et al., 1968). Within this thesis, the alpha level is kept constant at 0.05. For criteria showing normal distribution, paired-samples t-tests were ran to assess whether the mean difference between human and AI scores significantly differs from zero (Student, 1908). Cohen's d (C_d) was calculated as the mean difference divided by the standard deviation of the differences, providing a standardized effect size measure of the magnitude of the observed difference (Cohen, 2013; Field, 2024).

In cases where the distribution of paired differences deviated from normality, the Wilcoxon signed-rank test was employed. This nonparametric alternative ranks the absolute differences between paired observations, retains the sign of each difference, and evaluates whether the sum of Signed-Ranks significantly differs from the null hypothesis of symmetry around zero; it is thus appropriate when normality cannot be assumed (Conover, 1999; Wilcoxon, 1945).

Finally, correlations between human and AI scores were explored with both P_r and S_ρ . Pearson's correlation coefficient (P_r) quantifies the strength and direction of a linear relationship by standardizing the covariance between two continuous variables, with inference based on the t-distribution (Pearson, 1895). Spearman's rank correlation coefficient (S_ρ) was preferred when non-normality was observed as it is a nonparametric measure of rank correlation that does not rely on any specific distributional assumptions (Corder & Foreman, 2014; Field, 2024; Spearman, 1904).

CHAPTER IV

FINDINGS

4.1. University instructors' view on use of artificial intelligence in English writing assessment

This section presents survey findings from $N = 96$ university instructors across Turkey, addressing their adoption, perceptions, and expectations regarding artificial intelligence (AI) tools for assessing English writing.

4.1.1. Artificial intelligence adoption patterns

Survey results indicate widespread adoption of AI tools for writing assessment among instructors, with a substantial majority ($n = 79$, 82.29%) reporting prior usage of such tools, compared to a minority ($n = 17$, 17.71%) with no experience (see Table 4.1).

Table 4.1

Frequencies and Percentages for Survey Item 5

Have you previously used any AI-based tools (e.g., Grammarly, ChatGPT, Gemini) to assess student English writing?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
No	17	17.71	17.71
Yes	79	82.29	100.00

Analysis of tool preferences revealed a concentrated landscape dominated by mainstream applications. Among participants reporting AI usage, an open-ended question allowing multiple responses showed *ChatGPT by OpenAI* as the predominant choice ($n = 64$, 36.36%), followed closely by *Grammarly* ($n = 47$, 26.70%). Other tools trailed, including *Gemini*

by Google ($n = 14$, 7.95%), Turnitin ($n = 11$, 6.25%), and Claude by Anthropic ($n = 7$, 3.98%). Notably, the cumulative share of mentions reached 50% after just these top two tools (see Table 4.2).

A long list of other tools garnered minimal recognition ($n = 1$ per tool, 0.57%), including *Apple Intelligence*, *Canva*, *Codeium*, *ETS e-rater*, *Elicit*, *Gamma*, *Harpa AI*, *HuggingChat*, *InferKit*, *PenPAL*, *Twee*, and *Wordtune*.

These findings suggest a clear preference for widely available and general-purpose AI tools, particularly *ChatGPT* and *Grammarly*, in the context of writing assessment. Tools specifically designed for academic assessment or language education, such as *ETS e-rater* or *Turnitin*, were mentioned less frequently, which may point to a broader reliance on mainstream generative AI tools rather than domain-specific solutions.

Table 4.2
Frequencies and Percentages for Survey Item 6

Please specify which AI tool(s) you have used.	<i>n</i>	%	<i>Cumulative %</i>
Apple Intelligence	1	0.57	0.57
Canva	1	0.57	1.14
ChatGPT by OpenAI	64	36.36	37.50
Claude by Anthropic	7	3.98	41.48
Codeium	1	0.57	42.05
Copilot by Microsoft	6	3.41	45.45
DeepL	5	2.84	48.30
ETS e-rater	1	0.57	48.86
Elicit	1	0.57	49.43
Gamma	1	0.57	50.00

Continued on next page

Table 4.2*Frequencies and Percentages for Survey Item 6 (Continued)*

Please specify which AI tool(s) you have used.	<i>n</i>	%	<i>Cumulative %</i>
Gemini by Google	14	7.95	57.95
Grammarly	47	26.70	84.66
Harpa AI	1	0.57	85.23
HuggingChat	1	0.57	85.80
InferKit	1	0.57	86.36
PenPAL	1	0.57	86.93
Perplexity AI	5	2.84	89.77
QuillBot	5	2.84	92.61
Turnitin	11	6.25	98.86
Twee	1	0.57	99.43
Wordtune	1	0.57	100.00

Note. Disaggregated responses from open-ended item. Frequencies reflect total mentions across all participants.

4.1.2. Artificial intelligence employment strategies

To explore current usage behaviors, participants rated their present employment of AI tools in assessing English writing on a 5-point scale: from *not at all* to *very frequently*. Response distributions, detailed in Table 4.3, showed relatively even spread across the scale with a slight prevalence of higher ratings. Minimal usage (2 or lower) was reported by 34.38%. Cumulatively, 65.62% of respondents rated their usage at *sometimes* or more frequent, indicating moderate-to-high reliance on AI tools for assessment tasks.

Table 4.3*Frequencies and Percentages for Survey Item 7*

To what extent do you currently use AI tools (e.g., Grammarly, ChatGPT, Gemini) in assessing student English writing?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
Not at all	19	19.79	19.79
Rarely	14	14.58	34.38
Sometimes	20	20.83	55.21
Often	20	20.83	76.04
Very frequently	23	23.96	100.00

Participants' answers to the question about the primary function of AI tools used in assessing English writing were thematically coded into seven categories (see Table 4.4). This single-response item allowed participants either to choose one of the predefined options or to enter their own description.

Table 4.4*Frequencies and Percentages for Survey Item 8*

Which function of AI tools do you utilize the most in assessing student English writing?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
Content and argument analysis	13	13.40	13.40
Feedback on organization and structure	25	25.77	39.18
Grammar and syntax checking	23	23.71	62.89
Marking	3	3.09	65.98

Continued on next page

Table 4.4*Frequencies and Percentages for Survey Item 8 (Continued)*

Which function of AI tools do you utilize the most in assessing student English writing?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
Paraphrasing	1	1.03	67.01
Plagiarism detection	14	14.43	81.44
Preparing sample texts and answer keys	1	1.03	82.47
None (not using AI tools)	17	17.53	100.00

Note. Disaggregated responses from open-ended item. Frequencies reflect total mentions across all participants.

The most commonly cited function was *providing feedback on organization and structure*, with 25 mentions (25.77%), followed closely by *grammar and syntax checking*, cited 23 times (23.71%). These results suggest that educators predominantly rely on AI to support form-focused and discourse-level aspects of writing assessment.

Plagiarism detection also emerged as a relatively common use, receiving 14 mentions (14.43%), indicating a concern with academic integrity. Moreover, fewer participants reported using AI for *content and argument analysis* (13 mentions, 13.40%).

Less commonly cited functions included *marking* (3 mentions, 3.09%), *paraphrasing* (1 mention, 1.03%), and *preparing sample texts and answer keys* (1 mention, 1.03%). These lower frequencies may imply that such uses are either perceived as less pedagogically sound or less time-saving compared to more commonly adopted applications.

4.1.3. Perceptions of effectiveness of artificial intelligence

To investigate instructors' perceptions of AI effectiveness, five distinct assessment functions were evaluated: *grammar and syntax checking*, *plagiarism detection*, *content and argument*

analysis, feedback on organization and structure, and marking. Internal consistency across these dimensions was moderate to good ($C_\alpha = 0.675$).

Instructors reported the highest perceptions of effectiveness for mechanical, rule-based tasks. For grammar and syntax checking, 66.66% of respondents rated AI tools as effective, 2.08% judged them ineffective, and 31.25% were neutral, indicating strong trust in automated linguistic analysis (see Table 4.5).

Table 4.5

Frequencies and Percentages for Survey Item 9

In your opinion, how effective are AI tools in supporting the grammar and syntax checking aspect of assessing student English writing?	<i>n</i>	%	<i>Cumulative %</i>
[1] Very ineffective	2	2.08	2.08
[2] Ineffective	0	0.00	2.08
[3] Neutral/No opinion	30	31.25	33.33
[4] Effective	37	38.54	71.88
[5] Very effective	27	28.12	100.00

Perceptions of AI's effectiveness in plagiarism detection were somewhat lower yet remained positive: 33.34% affirmed its effectiveness, 9.38% dismissed it, and a majority of 57.29% were neutral, reflecting both appreciation for automated originality checks and caution about false positives (see Table 4.6).

Table 4.6*Frequencies and Percentages for Survey Item 10*

In your opinion, how effective are AI tools in supporting the plagiarism detection aspect of assessing student English writing?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
[1] Very ineffective	5	5.21	5.21
[2] Ineffective	4	4.17	9.38
[3] Neutral/No opinion	55	57.29	66.67
[4] Effective	22	22.92	89.58
[5] Very effective	10	10.42	100.00

By contrast, more interpretive tasks garnered more reserved perceptions. In content and argumentation analysis, only 40.62% deemed AI effective, while 20.83% considered it ineffective and 38.54% remained undecided, suggesting concerns about AI's capacity to assess rhetorical nuance (see Table 4.7).

Table 4.7*Frequencies and Percentages for Survey Item 11*

In your opinion, how effective are AI tools in supporting the content and argument analysis aspect of assessing student English writing?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
[1] Very ineffective	3	3.12	3.12
[2] Ineffective	17	17.71	20.83
[3] Neutral/No opinion	37	38.54	59.38
[4] Effective	32	33.33	92.71

Continued on next page

Table 4.7*Frequencies and Percentages for Survey Item 11 (Continued)*

In your opinion, how effective are AI tools in supporting the content and argument analysis aspect of assessing student English writing?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
[5] Very effective	7	7.29	100.00

Regarding feedback on organization and structure, 47.92% perceived AI guidance as effective, 11.46% found it ineffective, and 40.62% were neutral. This distribution highlights mixed views on AI's support for discourse coherence (see Table 4.8).

Table 4.8*Frequencies and Percentages for Survey Item 12*

In your opinion, how effective are AI tools in supporting the feedback on organization and structure aspect of assessing student English writing?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
[1] Very ineffective	2	2.08	2.08
[2] Ineffective	9	9.38	11.46
[3] Neutral/No opinion	39	40.62	52.08
[4] Effective	22	22.92	75.00
[5] Very effective	24	25.00	100.00

The lowest perception of effectiveness was observed in marking: only 13.54% rated AI scoring as effective, whereas 29.17% rated it ineffective and 57.29% were neutral (see Table 4.9). This pronounced hesitancy underscores ongoing concerns about fairness, transparency, and contextual sensitivity in automated scoring systems (J. Burstein & LaFlair, 2024).

Table 4.9*Frequencies and Percentages for Survey Item 13*

In your opinion, how effective are AI tools in supporting the marking aspect of assessing student English writing?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
[1] Very ineffective	10	10.42	10.42
[2] Ineffective	18	18.75	29.17
[3] Neutral/No opinion	55	57.29	86.46
[4] Effective	11	11.46	97.92
[5] Very effective	2	2.08	100.00

Overall, participants perceived AI as most effective for mechanical and objective aspects of writing assessment (grammar and syntax checking, and, to a lesser extent, plagiarism detection), while expressing caution about its role in interpretive and subjective tasks (content and argument analysis, feedback on organization and structure, and grading). These findings align with broader discussions on the complementary roles of human and AI evaluators rather than replacement.

4.1.4. Self-reported confidence, fairness, and pedagogical alignment of artificial intelligence

Participants rated their confidence in AI tools' ability to assess the quality of English writing, with neutrality emerging as the most frequent response (30.21%), yet a larger proportion of respondents expressed moderate to high confidence (42.71%) than low confidence (27.09%; see Table 4.10).

Table 4.10*Frequencies and Percentages for Survey Item 14*

How would you describe your confidence in AI tools' abilities to assess the quality of student English writing?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
[1] No confidence	10	10.42	10.42
[2] Low confidence	16	16.67	27.08
[3] Moderate confidence	29	30.21	57.29
[4] High confidence	25	26.04	83.33
[5] Full confidence	16	16.67	100.00

Similarly, perceptions of fairness and objectivity revealed a neutral midpoint as most common (38.54%), while positive evaluations (41.67%) exceeded negative ones (19.79%), indicating a generally favorable disposition toward AI-driven assessment (see Table 4.11).

Table 4.11*Frequencies and Percentages for Survey Item 15*

How do you describe the fairness and objectivity of AI tools in assessing student English writing?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
[1] Very unfair	5	5.21	5.21
[2] Somewhat unfair	14	14.58	19.79
[3] Neutral/No opinion	37	38.54	58.33
[4] Somewhat fair	28	29.17	87.50
[5] Very fair	12	12.50	100.00

The perceived alignment between AI tools and pedagogical objectives in English writing likewise showed a neutral plurality (30.21%), but more instructors agreed or strongly agreed (43.75%) than disagreed (26.04%) regarding this compatibility (see Table 4.12).

Table 4.12

Frequencies and Percentages for Survey Item 30

To what extent do you believe the use of AI aligns with pedagogical objectives for teaching English writing?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
[1] Not at all aligned	7	7.29	7.29
[2] Slightly aligned	18	18.75	26.04
[3] Neutral/No opinion	29	30.21	56.25
[4] Largely aligned	26	27.08	83.33
[5] Completely aligned	16	16.67	100.00

Across these three dimensions—confidence, fairness, and pedagogical alignment—a consistent pattern of cautious optimism emerged, supported by very high internal consistency with C_α of 0.921.

4.1.5. Perceived alignment with personal criteria and rubric-based assessment

When considering how well AI-generated assessments align with instructors’ own evaluative standards, responses were centered around neutrality or mild agreement, with 32.29% selecting the neutral midpoint and 33.33% indicating higher perceived alignment, while 34.38% expressed disagreement (see Table 4.13).

Table 4.13*Frequencies and Percentages for Survey Item 16*

To what extent do you feel AI tools align with your personal criteria for assessing student English writing?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
[1] Not at all aligned	11	11.46	11.46
[2] Slightly aligned	22	22.92	34.38
[3] Neutral/No opinion	31	32.29	66.67
[4] Largely aligned	15	15.62	82.29
[5] Completely aligned	17	17.71	100.00

In contrast, alignment perceptions improved when a standardized rubric was employed: although 35.42% remained neutral, an equal proportion of 35.42% agreed that AI assessments matched rubric criteria, and fewer (29.16%) disagreed (see Table 4.14). This shift suggests that the provision of standardized criteria may enhance educators' trust in AI-generated assessments, potentially due to the greater objectivity and consistency that rubrics can facilitate in both human and *machine* evaluations (Andrade, 2000; M. Wilson, 2006, 2018).

Table 4.14*Frequencies and Percentages for Survey Item 17*

When specific assessment criteria are provided, how closely do you find AI assessments align with your evaluations of student English writing based on the same rubric?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
[1] Not at all aligned	14	14.58	14.58

Continued on next page

Table 4.14*Frequencies and Percentages for Survey Item 17 (Continued)*

When specific assessment criteria are provided, how closely do you find AI assessments align with your evaluations of student English writing based on the same rubric?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
[2] Slightly aligned	14	14.58	29.17
[3] Neutral/No opinion	34	35.42	64.58
[4] Largely aligned	18	18.75	83.33
[5] Completely aligned	16	16.67	100.00

Overall, the excellent reliability between these two items ($C_\alpha = 0.927$) further indicates a coherent underlying construct of perceived alignment. The data indicate a cautiously optimistic stance among participants regarding the use of AI tools in writing assessment. While many remain neutral, the tendency for more positive responses (when rubrics are employed) points to the potential of structured assessment frameworks in mitigating skepticism toward AI integration.

4.1.6. Perceptions of ethical implications of using artificial intelligence

Participants indicated varying degrees of concern regarding ethical issues, such as data privacy and algorithmic bias, in AI-assisted English writing assessment (see Table 4.15). Responses reveal that while a small fraction of instructors (2.08%) expressed no concern at all and an additional 18.75% were only mildly concerned—together constituting 20.83% of the sample—approximately one third (33.33%) remained neutral on these matters.

Conversely, a substantial proportion reported elevated levels of apprehension: 30.21% described themselves as fairly concerned and 15.62% as very concerned, yielding a combined 45.83% of respondents whose concern exceeded neutrality. These results suggest that nearly

half of the surveyed instructors regard ethical considerations as significant when integrating AI into writing assessment, whereas about one fifth view such concerns as minimal.

Table 4.15

Frequencies and Percentages for Survey Item 29

How would you describe your stance on ethical issues (e.g., data privacy, bias) in AI-assisted student English writing assessment?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
[1] Not at all concerned	2	2.08	2.08
[2] Mildly concerned	18	18.75	20.83
[3] Neutral/No opinion	32	33.33	54.17
[4] Fairly concerned	29	30.21	84.37
[5] Very concerned	15	15.62	100.00

This pattern underscores a clear awareness among university instructors of the ethical dimensions inherent in deploying AI tools for evaluative purposes, aligning with prior research that highlights the critical need for ethical scrutiny in educational technologies (Holstein et al., 2019; Zawacki-Richter et al., 2019).

4.1.7. Role of artificial intelligence in the future

Responses to items assessing anticipated integration of AI into English writing assessment over the next 5 to 10 years reveal a strong consensus regarding its future ubiquity. As shown in Table 4.16, a majority of instructors endorsed the highest likelihood rating, with 57.29% selecting the top point on the scale and a further 28.12% indicating the second-highest level of confidence; only 1.04% expressed minimal expectation of such integration, yielding a combined optimistic response rate of 94.79%.

Table 4.16*Frequencies and Percentages for Survey Item 32*

How likely do you think it is that AI will become a common tool in assessing student English writing within the next 5–10 years?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
[1] Very unlikely	1	1.04	1.04
[2] Unlikely but possible	4	4.17	5.21
[3] Neutral/No opinion	9	9.38	14.58
[4] Somewhat likely	27	28.12	42.71
[5] Very likely	55	57.29	100.00

Similarly, when considering whether AI will assume a larger role in assessment processes (see Table 4.17), 61.46% of respondents chose the two most affirmative scale points (30.21% at the highest and 31.25% at the next), while outright rejection remained very low (2.08% at the lowest point).

Table 4.17*Frequencies and Percentages for Survey Item 33*

To what extent do you believe that AI could take on a larger role in assessing student English writing in the future?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
[1] Not at all	2	2.08	2.08
[2] To a limited extent	12	12.50	14.58
[3] Neutral/No opinion	23	23.96	38.54
[4] To a moderate extent	30	31.25	69.79

Continued on next page

Table 4.17*Frequencies and Percentages for Survey Item 33 (Continued)*

To what extent do you believe that AI could take on a larger role in assessing student English writing in the future?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
[5] Completely	29	30.21	100.00

Instructors' willingness to incorporate AI tools into regular assessment practice further underscores this forward-looking stance. As detailed in Table 4.18, 60.42% of participants indicated they would use AI routinely if the option were available, with 29.17% remaining undecided and only 10.42% expressing opposition.

Table 4.18*Frequencies and Percentages for Survey Item 35*

If you had the choice, would you incorporate AI regularly in your student English writing assessment process?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
[1] No	10	10.42	10.42
[2] Unsure	28	29.17	39.58
[3] Yes	58	60.42	100.00

The three items reflecting expectations of routine use, expanded role, and personal willingness exhibited strong internal consistency with C_α of 0.870, justifying their interpretation as a cohesive construct of future-focused acceptance.

4.1.8. Human oversight in artificial intelligence

Despite the optimism expressed in the future of AI in assessment, instructors clearly emphasize the necessity of human involvement in any automated assessment paradigm. As evidenced in Table 4.19, 91.67% of respondents rated human oversight as important or very important, with only 7.29% adopting a neutral stance and 1.04% minimally devaluing this role; no participants rated oversight as completely unimportant. These findings indicate that while educators are eager to embrace AI innovations, they consider human judgment indispensable in maintaining assessment quality and ethical standards.

Table 4.19
Frequencies and Percentages for Survey Item 34

How significant do you consider human oversight to be in AI-based student English writing assessment?	<i>n</i>	<i>%</i>	<i>Cumulative %</i>
[1] Not at all important	0	0.00	0.00
[2] Of little importance	1	1.04	1.04
[3] Neutral/No opinion	7	7.29	8.33
[4] Quite important	23	23.96	32.29
[5] Very important	65	67.71	100.00

Collectively, these results suggest that university instructors not only foresee AI becoming a routine component of writing assessment but also demonstrate readiness to engage with such tools—provided that meaningful human oversight is maintained. This prevailing enthusiasm likely reflects increasing exposure to AI-driven educational technologies and a growing appreciation for their potential to enhance current assessment practices.

4.2. Human and artificial intelligence scores for argumentative ESL essays

This section presents comprehensive findings from the collected scores across 60 argumentative English as a Second Language (ESL) essays, encompassing evaluations from 4 human and 25 AI raters. The analyses are alphabetically organized according to the IELTS Writing Task 2 Bands (IELTS-WT2B) criteria, followed by the total score analysis. For each of these analyses, the arithmetic mean of the scores were computed on a per-essay level. These aggregate scores serve as the basis for all subsequent human-AI analysis.

4.2.1. Human scores analysis

Human raters demonstrated high inter-rater agreement, with Krippendorff's alpha (K_α) values exceeding the 0.7 threshold across every dimension—indicating above chance levels (per Krippendorff, 2019, pp. 356–357). Consistent with this robust reliability value, the intraclass correlation coefficient (ICC_{2k}) also remained above 0.94 for all criteria, indicating excellent consistency among raters and further justifying reliance on mean scores for further analyses (Koo & Li, 2016).

4.2.1.1. IELTS-WT2B criterion: Coherence and cohesion

Human raters demonstrated substantial agreement in assigning Coherence and Cohesion (CC) scores, with K_α of 0.714 (95% CI [0.587, 0.810]). An ICC_{2k} of 0.956 ($p < 0.01$) confirms excellent consistency in how human raters distributed their scores for the CC criterion.

The scores ranged from a minimum of 4.25 to a maximum of 8.50, with an overall mean of 6.02 ($SE = 0.14$; 95% CI [5.75, 6.31]). The standard deviation of 1.09 and variance of 1.19 indicate moderate dispersion around the mean. The distribution exhibits slight positive skewness (0.54), suggesting a modest tail towards higher scores, while the kurtosis of -0.233 reflects a distribution marginally flatter than normal (see Table 4.20).

Table 4.20

Descriptives of Aggregated Human Scores for the IELTS-WT2B “Coherence and Cohesion” Criterion

Descriptive	Statistic	Std. Error	95% Confidence Interval [*]	
			Lower	Upper
<i>N</i>	60			
Minimum	4.25			
Maximum	8.50			
Mean	6.02	0.14	5.75	6.31
Std. Deviation	1.09		0.89	1.24
Variance	1.19		0.78	1.54
Skewness	0.54	0.30	0.15	0.93
Kurtosis	−0.23	0.60	−0.95	0.73

^{*} Calculated via bootstrap resampling at 1000 iterations.

4.2.1.2. IELTS-WT2B criterion: Grammatical range and accuracy

Human raters showed high agreement when grading Grammatical Range and Accuracy, with K_{α} of 0.802 (95% CI [0.718, 0.854]). An ICC_{2k} of 0.955 ($p < 0.01$) indicates excellent consistency among human raters in distributing these scores.

The scores ranged from a minimum of 3.75 to a maximum of 8.50, with an overall mean of 5.83 ($SE = 0.16$; 95% CI [5.50, 6.17]). The standard deviation of 1.23 and variance of 1.50 denote moderate dispersion around the mean. The distribution shows slight positive skewness (0.28), indicating a small tail toward higher scores, while the kurtosis of −0.72 reflects a marginally flatter than normal peak (see Table 4.21).

Table 4.21

Descriptives of Aggregated Human Scores for the IELTS-WT2B “Grammatical Range and Accuracy” Criterion

Descriptive	Statistic	Std. Error	95% Confidence Interval [*]	
			Lower	Upper
<i>N</i>	60			
Minimum	3.75			
Maximum	8.50			
Mean	5.83	0.16	5.50	6.17
Std. Deviation	1.23		1.02	1.40
Variance	1.50		1.03	1.96
Skewness	0.28	0.31	−0.13	0.67
Kurtosis	−0.72	0.61	−1.19	0.00

^{*} Calculated via bootstrap resampling at 1000 iterations.

4.2.1.3. IELTS-WT2B criterion: Lexical resource

For Lexical Resource, human raters attained K_α of 0.758 (95% CI [0.661, 0.824]), confirming strong agreement. An ICC_{2k} of 0.955 ($p < 0.01$) further confirms the consistency of the distribution of these scores.

The scores ranged from a minimum of 4.00 to a maximum of 8.50, with an overall mean of 6.16 ($SE = 0.16$; 95% CI [5.87, 6.48]). The standard deviation of 1.21 and variance of 1.46 denote moderate variability around the mean. The distribution exhibits low positive skewness (0.16), indicating near symmetry with a slight tail toward higher scores, while the kurtosis of −0.92 suggests a flatter than normal distribution with lighter tails (see Table 4.22).

Table 4.22

Descriptives of Aggregated Human Scores for the IELTS-WT2B “Lexical Resource” Criterion

Descriptive	Statistic	Std. Error	95% Confidence Interval [*]	
			Lower	Upper
<i>N</i>	60			
Minimum	4.00			
Maximum	8.50			
Mean	6.16	0.16	5.87	6.48
Std. Deviation	1.21		1.02	1.35
Variance	1.46		1.05	1.81
Skewness	0.16	0.31	−0.26	0.57
Kurtosis	−0.92	0.61	−1.27	−0.41

* Calculated via bootstrap resampling at 1000 iterations.

4.2.1.4. IELTS-WT2B criterion: Task response

Task Response (TR) scores exhibited the highest inter-rater agreement among the human raters, with K_α of 0.860 (95% CI [0.801, 0.901]). An ICC_{2k} of 0.946 ($p < 0.01$) confirms that human raters consistently distributed TR scores within a narrow band.

The scores ranged from a minimum of 4.25 to a maximum of 9.00, with an overall mean of 6.55 ($SE = 0.14$; 95% CI [6.28, 6.86]). The standard deviation of 1.12 and variance of 1.25 indicate moderate dispersion around the mean. The distribution exhibits near symmetry, with slight positive skewness (0.13), while the kurtosis of −0.93 reflects a platykurtic shape with relatively light tails (see Table 4.23).

Table 4.23*Descriptives of Aggregated Human Scores for the IELTS-WT2B “Task Response” Criterion*

Descriptive	Statistic	Std. Error	95% Confidence Interval [*]	
			Lower	Upper
<i>N</i>	60			
Minimum	4.25			
Maximum	9.00			
Mean	6.55	0.14	6.28	6.86
Std. Deviation	1.12		0.96	1.25
Variance	1.25		0.92	1.57
Skewness	0.13	0.31	−0.29	0.57
Kurtosis	−0.93	0.61	−1.33	−0.36

^{*} Calculated via bootstrap resampling at 1000 iterations.

4.2.1.5. IELTS-WT2B criteria: The total score

When considering the total score (sum of all four criteria, maximum 36), human raters again demonstrated exceptional agreement with K_{α} of 0.891 (95% CI [0.844, 0.922]). An ICC_{2k} of 0.964 ($p < 0.01$) further underlines that human raters applied their scoring criteria in a highly consistent manner across essays.

The total scores spanned from a minimum of 17.50 to a maximum of 32.50. The mean score was 24.56 ($SE = 0.52$; 95% CI [23.55, 25.65]), indicating a central tendency near the mid-scale range. Score variability was moderate, with a standard deviation of 4.01 and a corresponding variance of 16.07. The distribution demonstrated slight positive skewness (0.28), suggesting a mild right-tail and platykurtic behavior with a kurtosis value of −0.59 (see Table 4.24).

Table 4.24*Descriptives of Aggregated Human Scores for the sum of IELTS-WT2B Criteria*

Descriptive	Statistic	Std. Error	95% Confidence Interval [*]	
			Lower	Upper
<i>N</i>	60			
Minimum	17.50			
Maximum	32.50			
Mean	24.56	0.52	23.55	25.65
Std. Deviation	4.01		3.35	4.58
Variance	16.07		11.20	21.01
Skewness	0.28	0.31	−0.13	0.66
Kurtosis	−0.59	0.61	−1.07	0.09

^{*} Calculated via bootstrap resampling at 1000 iterations.

4.2.2. Artificial intelligence scores analysis

The analysis of AI scores revealed greater variability and lower inter-rater agreement among the AI raters compared to human raters. Across all dimensions, Krippendorff's alpha (K_α) values reflect moderate to low levels of agreement among the 25 large language models (LLMs) operating under a zero-shot prompting protocol. Despite this variability, the intraclass correlation coefficient (ICC_{2k}) values suggest that the AI scores still yield stable rank orderings, pointing to internal consistency.

4.2.2.1. IELTS-WT2B criterion: Coherence and cohesion

AI scores for the Coherence and Cohesion criterion attained K_α of 0.468 (95% CI [0.353, 0.545]), indicating only moderate agreement among raters. An ICC_{2k} of 0.956 ($p < 0.01$)

indicates that, despite variability in individual grading outputs, the aggregate of AI scores yields a consistent rank ordering when averaged across all 25 AI raters.

The aggregated AI scores ranged from a minimum of 3.84 to a maximum of 6.88, with an overall mean of 5.57 ($SE = 0.80$; 95% CI [5.41, 5.72]). The standard deviation of 0.62 reflects moderate dispersion around the mean, and the corresponding variance of 0.39 further quantifies this variability. The distribution exhibits slight negative skewness at -0.20 , indicating a marginal tail toward lower scores, while the kurtosis of 0.27 suggests a distribution that is close to mesokurtic in terms of peakedness and tail weight (see Table 4.25).

Table 4.25

Descriptives of Aggregated AI Scores for the IELTS-WT2B “Coherence and Cohesion” Criterion

Descriptive	Statistic	Std. Error	95% Confidence Interval [*]	
			Lower	Upper
<i>N</i>	60			
Minimum	3.84			
Maximum	6.88			
Mean	5.57	0.80	5.41	5.72
Std. Deviation	0.62		0.50	0.73
Variance	0.39		0.25	0.54
Skewness	-0.20	0.31	-0.72	0.45
Kurtosis	0.27	0.61	-0.76	1.41

^{*} Calculated via bootstrap resampling at 1000 iterations.

4.2.2.2. IELTS-WT2B criterion: Grammatical range and accuracy

For Grammatical Range and Accuracy, AI raters achieved K_α of 0.456 (95% CI [0.358, 0.532]), again indicating moderate inter-rater agreement. An ICC_{2k} of 0.955 ($p < 0.01$) confirms that, although individual rater judgments differ, their aggregated distribution maintains a stable ordering.

The aggregated AI scores ranged from a minimum of 4.20 to a maximum of 6.84, with an overall mean of 5.28 ($SE = 0.08$; 95% CI [5.13, 5.42]). The standard deviation of 0.59 indicates moderate dispersion around the mean, and the variance of 0.35 further quantifies this variability. The distribution displays a modest positive skewness of 0.39, suggesting a slight tail toward higher scores, while the kurtosis of 0.09 implies a distribution close to mesokurtic in terms of peakedness and tail weight (see Table 4.26).

Table 4.26

Descriptives of Aggregated AI Scores for the IELTS-WT2B “Grammatical Range and Accuracy” Criterion

Descriptive	Statistic	Std. Error	95% Confidence Interval [*]	
			Lower	Upper
<i>N</i>	60			
Minimum	4.20			
Maximum	6.84			
Mean	5.28	0.08	5.13	5.42
Std. Deviation	0.59		0.47	0.69
Variance	0.35		0.22	0.48
Skewness	0.39	0.31	−0.19	0.81
Kurtosis	0.09	0.61	−0.83	1.13

^{*} Calculated via bootstrap resampling at 1000 iterations.

4.2.2.3. IELTS-WT2B criterion: Lexical resource

AI raters' Lexical Resource scores yielded K_α of 0.463 (95% CI [0.358, 0.535]), denoting moderate agreement. An ICC_{2k} of 0.955 ($p < 0.01$) shows that, despite inter-rater differences, the collective AI score distribution is internally consistent when averaged.

The aggregated AI scores ranged from a minimum of 4.48 to a maximum of 7.16, with an overall mean of 5.59 ($SE = 0.07$; 95% CI [5.45, 5.72]). The standard deviation of 0.56 indicates moderate dispersion around the mean, and the variance of 0.31 further characterizes this variability. The distribution exhibits slight positive skewness at 0.20, suggesting a modest tail toward higher scores, while the kurtosis of 0.13 implies a distribution that is approximately mesokurtic with respect to peakedness and tail weight (see Table 4.27).

Table 4.27

Descriptives of Aggregated AI Scores for the IELTS-WT2B “Lexical Resource” Criterion

Descriptive	Statistic	Std. Error	95% Confidence Interval*	
			Lower	Upper
<i>N</i>	60			
Minimum	4.48			
Maximum	7.16			
Mean	5.59	0.07	5.45	5.72
Std. Deviation	0.56		0.45	0.65
Variance	0.31		1.05	1.81
Skewness	0.20	0.31	−0.47	0.71
Kurtosis	0.13	0.61	−0.90	1.29

* Calculated via bootstrap resampling at 1000 iterations.

4.2.2.4. IELTS-WT2B criterion: Task response

AI scores for the Task Response criterion exhibited the lowest inter-rater reliability, with K_α of 0.421 (95% CI [0.307, 0.512]). An ICC_{2k} of 0.946 ($p < 0.01$) suggests that, although the raters differ more in this dimension than in others, their averaged rank ordering remains reasonably stable.

The aggregated AI scores ranged from a minimum of 3.92 to a maximum of 6.96, with an overall mean of 5.64 ($SE = 0.08$; 95% CI [5.49, 5.79]). The standard deviation of 0.60 denotes moderate dispersion around the mean, and the variance of 0.36 further quantifies this variability. The distribution exhibits a modest negative skewness of -0.36 , indicating a slight tail toward lower scores, while the kurtosis of 0.85 suggests a somewhat more peaked distribution with heavier central mass than a normal distribution (see Table 4.28).

Table 4.28

Descriptives of Aggregated AI Scores for the IELTS-WT2B “Task Response” Criterion

Descriptive	Statistic	Std. Error	95% Confidence Interval*	
			Lower	Upper
<i>N</i>	60			
Minimum	3.92			
Maximum	6.96			
Mean	5.64	0.08	5.49	5.79
Std. Deviation	0.60		0.47	0.72
Variance	0.36		0.22	0.52
Skewness	-0.36	0.31	-0.94	0.51
Kurtosis	0.85	0.61	-0.67	2.05

* Calculated via bootstrap resampling at 1000 iterations.

4.2.2.5. IELTS-WT2B criteria: The total score

Considering the total of AI scores (sum of all four criteria, maximum 36), K_α of 0.508 (95% CI [0.402, 0.581]) indicates moderate inter-rater agreement. An ICC_{2k} of 0.964 ($p < 0.01$) confirms that, despite variability among individual raters, the aggregated AI score ranking is consistent when averaged across raters.

The total scores across the 60 essays, as graded by the ensemble of 25 AI raters, ranged from a minimum of 16.76 to a maximum of 27.44. The observed mean was 22.09 ($SE = 0.30$; 95% CI [21.51, 22.63]), indicating that the central tendency lay slightly above the midpoint of the scoring scale. Variability around this mean was modest, with a standard deviation of 2.32 and a variance of 5.38. The distribution of scores was approximately symmetrical, as reflected by a skewness of 0.05, and exhibited near-mesokurtic characteristics with a kurtosis of 0.07 (see Table 4.29).

Table 4.29

Descriptives of Aggregated AI Scores for the Sum of IELTS-WT2B Criteria

Descriptive	Statistic	Std. Error	95% Confidence Interval [*]	
			Lower	Upper
<i>N</i>	60			
Minimum	16.76			
Maximum	27.44			
Mean	22.09	0.30	21.51	22.63
Std. Deviation	2.32		1.86	2.72
Variance	5.38		3.47	7.39
Skewness	0.05	0.31	−0.49	0.57
Kurtosis	0.07	0.61	−0.67	0.94

^{*} Calculated via bootstrap resampling at 1000 iterations.

4.2.3. Drawing comparisons between human and artificial intelligence scores

As shown in Table 4.31, human and AI scores yielded a range of Pearson’s correlation coefficient (P_r) values, all reaching statistical significance at $p < 0.001$. It is important to note, however, that these correlations were computed using only the arithmetic mean values of human and AI scores for each of the IELTS-WT2B criterion and for the total score; the “IELTS-rounded mean” rows are included solely for completeness and were not used in the analyses.

Prior to conducting the correlational analysis, the distribution of paired difference scores were examined $\Delta_i = \bar{x}(AI_i) - \bar{x}(\text{Human}_i)$ for normality via the Shapiro-Wilk test (Shapiro & Wilk, 1965). Across all criteria, with the exception of the IELTS-WT2B criterion Lexical Resource, the test yielded W values of at least 0.959 with corresponding p values exceeding 0.05—indicating no significant departure from normality (see Table 4.30). Consequently, parametric tests—namely the paired-samples t-test for systematic bias and Cohen’s d (C_d) for effect-size estimation—were conducted following the procedures defined in Student (1908) and Cohen (2013). For the LR criterion, however, the Wilcoxon signed-rank test (Pratt, 1959; Wilcoxon, 1945) was applied instead of the paired t-test due to non-normality. In addition, only Spearman’s rank correlation coefficient (S_ρ) was employed to assess the association for this criterion.

Table 4.30

Shapiro-Wilk Normality Test Results for Paired Difference Scores Across the IELTS-WT2B Criteria

IELTS-WT2B Criterion	Shapiro-Wilk		
	Statistic (W)	df	Significance (p)
Coherence and Cohesion	0.980	60	0.409
Grammatical Range and Accuracy	0.966	60	0.098

Continued on next page

Table 4.30

Shapiro-Wilk Normality Test Results for Paired Difference Scores Across the IELTS-WT2B Criteria (Continued)

Lexical Resource [†]	0.959	60	0.041
Task Response	0.967	60	0.104
Total Score	0.983	60	0.563

[†] Non-normal distribution.

Although P_r was selected as the primary measure of linear correlation for those criteria meeting normality assumptions—given its greater statistical power under such conditions (Field, 2024; Kendall et al., 1973)— S_ρ values are reported for completeness in Table 4.31, and were the sole correlation coefficient applied to the Lexical Resource criterion. Scatter plots illustrating the relationships for each individual criterion and for the total score are provided in Figures 4.1, 4.2, 4.3, 4.4, and 4.5.

Table 4.31

Correlations Between Aggregated Human and AI Scores

IELTS-WT2B Criterion		Pearson's r (P_r)	Spearman's rho (S_ρ)
Coherence and Cohesion	Arithmetic mean ($\bar{x}$)	0.513 [*]	0.516 [*]
	IELTS-rounded mean ^a	0.452 [*]	0.454 [*]
Grammatical Range and Accuracy	Arithmetic mean ($\bar{x}$)	0.580 [*]	0.560 [*]
	IELTS-defined mean ^a	0.513 [*]	0.510 [*]

Continued on next page

Table 4.31*Correlations Between Aggregated Human and AI Scores (Continued)*

Lexical Resource [†]	Arithmetic mean ($\bar{x}$)	0.593 [*]	0.553 [*]
	IELTS-defined mean ^a	0.602 [*]	0.573 [*]
Task Response	Arithmetic mean ($\bar{x}$)	0.560 [*]	0.576 [*]
	IELTS-defined mean ^a	0.573 [*]	0.504 [*]
Total Score	Arithmetic mean ($\bar{x}$)	0.654 [*]	0.644 [*]
	IELTS-defined mean ^a	0.672 [*]	0.673 [*]

^{*} $p < 0.001$, two-tailed.

[†] Non-normal distribution.

^a The total score is calculated as the arithmetic mean of the four criteria, then rounded to the nearest half-number (IELTS, 2025).

4.2.3.1. IELTS-WT2B criterion: Coherence and cohesion

Descriptive analysis of the paired difference scores ($\Delta_{CC} = AI_{CC} - Human_{CC}$) for the Coherence and Cohesion (CC) criterion revealed a mean bias of -0.45 (95% CI $[-0.69, -0.21]$), indicating that, on average, AI raters awarded CC scores ≈ 0.5 points lower than human raters. The variability of this bias ($SD = 0.94$) suggests that most discrepancies fell within approximately one point of the human CC scores. The observed range (-3.06 to 1.55) further highlights that some AI raters under-scored by up to three points, while others overscored by up to one and a half points relative to the human CC scores.

To assess whether this mean bias significantly differed from zero, a paired-samples t-test ($N = 60$) was conducted. The difference in CC scores was significantly negative, paired $t(59) = -3.71$ ($p < 0.001$). The standardized effect size was moderate ($C_d = -0.479$, 95% CI $[-0.745, -0.210]$), indicating a moderate underestimation bias by the AI raters.

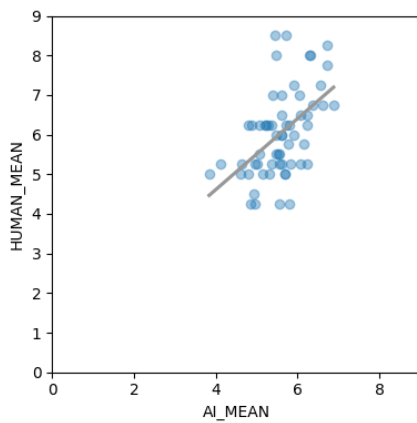


Figure 4.1
Scatter Plot Comparing Mean Human and AI “Coherence and Cohesion” Scores

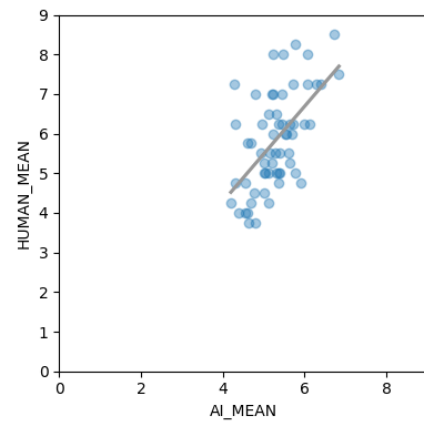


Figure 4.2
Scatter Plot Comparing Mean Human and AI “Grammatical Range and Accuracy” Scores

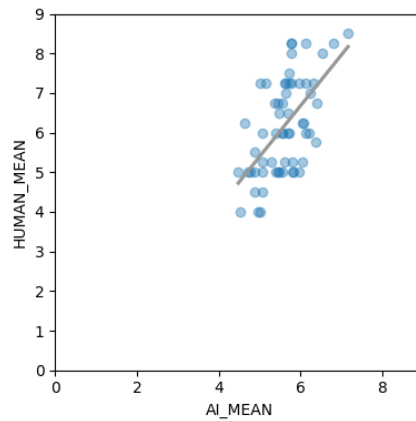


Figure 4.3
Scatter Plot Comparing Mean Human and AI “Lexical Resource” Scores

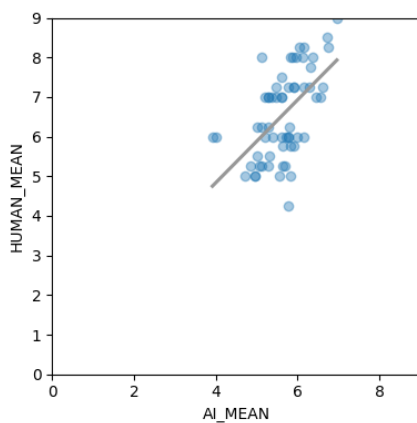


Figure 4.4
Scatter Plot Comparing Mean Human and AI “Task Response” Scores

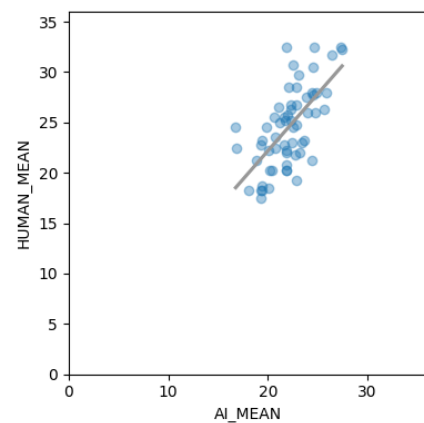


Figure 4.5
Scatter Plot Comparing Mean Human and AI Total Scores

Furthermore, AI and human scores were moderately correlated, with $P_r = 0.513$ ($p < 0.001$), representing the lowest correlation coefficient among the IELTS-WT2B criteria.

When examining individual AI raters, *OpenAI o3-mini* exhibited the highest alignment with mean human gradings ($P_r = 0.581$), whereas *Google Gemini 1.5 Flash-8B* showed the lowest alignment ($P_r = 0.217$). This wide range (0.217 to 0.581) indicates substantial variability among AI raters in reproducing mean human judgments on the CC criterion.

4.2.3.2. IELTS-WT2B criterion: Grammatical range and accuracy

Descriptive analysis of the paired difference scores ($\Delta_{GRA} = AI_{GRA} - Human_{GRA}$) for the Grammatical Range and Accuracy (GRA) criterion revealed a mean bias of -0.542 (95% CI $[-0.802, -0.282]$), indicating that, on average, AI raters awarded GRA scores ≈ 0.5 points lower than human raters. The variability of this bias ($SD = 1.006$) suggests that most discrepancies fell within approximately one point of the human GRA scores. The observed range (-2.970 to 1.170) further highlights that some AI raters under-scored by nearly three points, while others overscored by just over one point relative to human judgments.

To assess whether this mean bias significantly differed from zero, a paired-samples t-test ($N = 60$) was conducted. The difference in GRA scores was significantly negative, paired $t(59) = -4.175$ ($p < 0.001$), demonstrating a consistent underestimation by the AI raters. The standardized effect size was moderate ($C_d = -0.539$, 95% CI $[-0.808, -0.266]$), indicating a meaningful underestimation bias in the AI assessments. Furthermore, AI and human scores showed a moderate level of correlation overall, with $P_r = 0.580$ ($p < 0.001$).

When examining individual AI raters, *Anthropic Claude 3.5 Sonnet (New)* exhibited the highest alignment with mean human gradings ($P_r = 0.574$), whereas *Google Gemini 2.0 Flash* showed the lowest alignment ($P_r = 0.159$). This wide range (0.159 to 0.574) shows substantial variability among AI raters in reproducing human judgments on the GRA criterion.

4.2.3.3. IELTS-WT2B criterion: Lexical resource

Descriptive analysis of the paired difference scores ($\Delta_{LR} = AI_{LR} - Human_{LR}$) for the Lexical Resource (LR) criterion revealed a mean bias of -0.564 (95% CI $[-0.819, -0.310]$), indicating that, on average, AI raters awarded LR scores ≈ 0.5 points lower than human raters. The variability of this bias ($SD = 0.984$) suggests that most discrepancies fell within approximately one point. The observed range (-2.490 to 1.000) highlights that some AI raters under-scored by nearly two and a half points, while others slightly overscored by up to one point relative to human scores.

A Shapiro-Wilk test was conducted to assess normality of the difference scores, yielding $W = 0.959$ ($p = 0.041$), which indicates a departure from normality. Consequently, nonparametric methods were employed instead of P_r , paired t-tests, and Cohen's d .

The Wilcoxon Signed-Ranks test ($N = 60$) based on negative ranks showed a significant negative bias in AI scoring, $Z = -3.633$, with 20 negative ranks and 40 positive ranks (asymptotic $p < 0.001$). This confirms that AI raters systematically under-scored the LR criterion compared to human raters.

Spearman's rank correlation coefficient (S_ρ) demonstrated a moderate positive association between AI and human LR scores, $S_\rho = 0.553$ ($p < 0.001$), suggesting reasonable but imperfect alignment. Among individual AI raters, *OpenAI o3-mini* achieved the highest correlation with human ratings ($S_\rho = 0.581$), whereas *Google Gemini 1.5 Flash-8B* exhibited the lowest ($S_\rho = 0.094$), illustrating substantial variability in rater performance on the LR criterion.

4.2.3.4. IELTS-WT2B criterion: Task Response

Descriptive analysis of the paired difference scores ($\Delta_{TR} = AI_{TR} - Human_{TR}$) for the Task Response (TR) criterion revealed a mean bias of -0.910 (95% CI $[-1.149, -0.671]$), indicating that, on average, AI raters awarded TR scores ≈ 1 point lower than human raters. The variability of this bias ($SD = 0.926$) suggests that most discrepancies fell within approximately one point of the human TR scores. The observed range (-2.880 to 1.510) further highlights that

some AI raters under-scored by up to nearly three points, while others overscored by up to one and a half points relative to the human TR scores.

To assess whether this mean bias significantly differed from zero, a paired-samples t-test ($N = 60$) was conducted. The difference in TR scores was highly significant, paired $t(59) = -7.615$ ($p < 0.001$), demonstrating a robust underestimation bias by the AI raters. The standardized effect size was large ($C_d = -0.983$, 95% CI $[-1.289, -0.671]$), indicating a substantial practical difference between AI and human ratings on the TR criterion.

Furthermore, AI and human scores were moderately correlated, with $P_r = 0.560$ ($p < 0.001$), reflecting a moderate alignment overall. Examination of individual AI raters revealed that *OpenAI o3-mini* exhibited the highest correspondence with human gradings ($P_r = 0.657$), whereas *Google Gemini 1.5 Flash-8B* showed the lowest correspondence ($P_r = 0.170$). This wide range (0.170 to 0.657) displays considerable variability among raters in reproducing human judgments on the TR criterion.

4.2.3.5. IELTS-WT2B criteria: The total score

Descriptive analysis of the paired difference scores ($\Delta = \text{AI} - \text{Human}$) for the total score (sum of four criteria, maximum of 36) revealed a mean bias of -2.466 (95% CI $[-3.254, -1.679]$), indicating that, on average, AI raters' total scores turn out ≈ 2.5 points lower than human raters. The variability of this bias ($SD = 3.048$) suggests that most discrepancies fell within approximately three points of the human total scores. The observed range (-10.700 to 3.590) further highlights that the sum of scores assigned by certain AI raters underestimated human reference scores by as much as ten and a half points, while AI other raters overestimated them by up to three and a half points.

To assess whether this mean bias significantly differed from zero, a paired-samples t-test ($N = 60$) was conducted. The difference in total scores was significantly negative, paired $t(59) = -6.268$ ($p < 0.001$). The standardized effect size was large ($C_d = -0.809$, 95% CI $[-1.099, -0.515]$), indicating a substantial underestimation bias by the AI raters. Furthermore, AI and human total scores were strongly correlated, with $P_r = 0.654$ ($p < 0.001$), suggesting considerable overall alignment between the two raters.

When examining individual AI raters, *OpenAI o3-mini* exhibited the highest alignment with mean human gradings ($P_r = 0.750$), whereas *Google Gemini 1.5 Flash-8B* showed the lowest alignment ($P_r = 0.178$). This wide range (0.178 to 0.750) demonstrates substantial variability among AI raters in reproducing human holistic scoring on the IELTS-WT2B.

4.2.3.6. Top 5 artificial intelligence raters by correlation coefficients

The AI rater *OpenAI o3-mini* demonstrated consistently strong performance, in leading rank across three of the IELTS-WT2B criteria as well as the total score (see Table 4.32). Another LLM developed by OpenAI, the *GPT-4o*, also showed considerably high performance, being in the top 5 in two of the IELTS-WT2B criteria as well as the total score.

These results indicate that the *OpenAI o3-mini* and *GPT-4o* not only align closely with mean human gradings on each individual criterion—especially in a theoretically high-impact criterion such as Task Response—but also attain high overall correlation with human total scores. Such consistent top-ranked performance showcases these models’ relative robustness in approximating aggregate human judgments across multiple dimensions of writing assessment.

Table 4.32
Top 5 AI Raters by Pearson’s Correlation Coefficients

Rank	Coherence and Cohesion	Grammatical Range and Accuracy	Lexical Resource†	Task Response	Total Score
1	OpenAI o3-mini $P_r = 0.581$	Anthropic Claude 3.5 Sonnet (New) $P_r = 0.574$	OpenAI o3-mini $S_\rho = 0.581$	OpenAI o3-mini $P_r = 0.657$	OpenAI o3-mini $P_r = 0.750$
2	OpenAI o1 $P_r = 0.570$	Anthropic Claude 3.7 Sonnet (Thinking Disabled) $P_r = 0.570$	OpenAI o3 $S_\rho = 0.562$	Google Gemini 2.0 Flash-Lite $P_r = 0.609$	Anthropic Claude 3.5 Sonnet (New) $P_r = 0.651$
3	Google Gemini 2.5 Pro Experimental 03-25 $P_r = 0.535$	Anthropic Claude 3.7 Sonnet (Thinking Enabled) $P_r = 0.554$	OpenAI GPT-4.1 $S_\rho = 0.550$	OpenAI GPT-4.1 mini $P_r = 0.587$	OpenAI GPT-4o $P_r = 0.633$
4	DeepSeek V3-0324 $P_r = 0.493$	OpenAI GPT-4o $P_r = 0.551$	Google Gemini 2.5 Flash Preview 04-17 $S_\rho = 0.545$	DeepSeek V3-0324 $P_r = 0.567$	OpenAI GPT-4.1 mini $P_r = 0.618$

Continued on next page

Table 4.32
Top 5 AI Raters by Pearson’s Correlation Coefficients (Continued)

Rank	Coherence and Cohesion	Grammatical Range and Accuracy	Lexical Resource [†]	Task Response	Total Score
5	Anthropic Claude	Anthropic Claude	Google Gemini 1.5	OpenAI GPT-4o	OpenAI o1
	3.5 Sonnet (Old)	3.5 Haiku	Flash	$P_r = 0.557$	$P_r = 0.610$
	$P_r = 0.489$	$P_r = 0.525$	$S_\rho = 0.510$		

[†] Non-normality observed. Spearman’s rho is listed instead.

CHAPTER V

CONCLUSION, DISCUSSION, AND SUGGESTIONS

5.1. Discussion

5.1.1. Adoption patterns across contexts

5.1.1.1. Historical trajectory and the scope of current usage

Early automated essay scoring (AES) systems such as Page's (1966) Project Essay Grade (PEG) leveraged surface-level proxies (e.g., word counts, syntactic markers) and laid the groundwork for statistical and rule-based enhancements in the 1980s–1990, such as latent semantic analysis (LSA), e-rater, and IntelliMetric (S. Elliot, 2003; Landauer et al., 1998). The literature reveals a gradual shift toward transformer-based automated writing evaluation (AWE) platforms that integrate large language models (LLMs) for richer representations and hybrid workflows—artificial intelligence (AI) assistance for humans (Taghipour & Ng, 2016; Xiao et al., 2025).

In parallel, the findings from the survey of 96 Turkish university instructors reveals a high prevalence of AI tools in English writing assessment, with 82.30% reporting prior use. ChatGPT (36.40%) and Grammarly (26.70%) dominate mentions, while specialized educational systems (e.g., ETS e-rater and Turnitin) remain marginal (0.60% per). This concentration on general-purpose LLMs reflects the importance of both accessibility and perceived versatility.

5.1.1.2. Frequency and functions of use

Literature on AWE emphasizes its dual role in providing rapid scoring and diagnostic feedback, particularly for grammar, mechanics, and organization when scaffolded by teacher mediation (J. Burstein, 2012; Stevenson & Phakiti, 2019). Consistently, 65.60% of instructors stated use AI tools at least occasionally, principally for feedback on organization and structure (25.80%) and grammar and syntax checking (23.70%). Notably, only 3.10% employ AI tools primarily for marking, indicating unwillingness to delegate summative evaluation entirely to AI.

5.1.1.3. Perceived effectiveness and trust

Studies report moderate positive pedagogical impacts of AWE when combined with peer or teacher support (J. Wilson & Roscoe, 2020), yet concerns persist regarding superficiality of feedback and threat to self-regulation (Deane, 2013; Ranalli, 2021). Survey respondents mirror these ambivalences: while 66.70% rate AI tools' ability to check grammar and syntax as effective, only 13.50% trust AI grading, and 57.30% remain neutral. Confidence in AI assessment is moderate to high for 42.70% of instructors, but 45.80% express fairly strong ethical concerns about bias and fairness. This cautious endorsement echoes literature critiques on construct validity and algorithmic bias (Bender et al., 2021; Perelman, 2014).

5.1.2. Pedagogical applications and the synergy between human and artificial intelligence

5.1.2.1. Hybrid models in practice

Contemporary scholarship advocates hybrid raters—AI tools flag inconsistencies or surface errors, human raters evaluate higher-order skills—to enhance reliability and explainability (Attali et al., 2010; Xiao et al., 2025). The correlational analysis of human and AI scores confirms that, although AI tools achieve high intraclass correlation coefficient (ICC_{2k}) values (above 0.94), they under-score essays relative to human raters by approximately 0.5 to 1.0 points depending on the IELTS Writing Task 2 Bands (IELTS-WT2B) criterion, with largest bias in Task Response (≈ -1 point) and the total score (≈ -2.5 points). This systematic under-scoring represents the necessity of human calibration to mitigate mean bias and preserve pedagogical integrity.

5.1.2.2. Criterion-specific insights

Human raters demonstrated excellent reliability, per Krippendorff's alpha (K_α), across the IELTS-WT2B criteria ($K_\alpha > 0.70$), particularly for Task Response ($K_\alpha = 0.86$) and the total score ($K_\alpha = 0.89$). AI raters, by contrast, achieved only moderate K_α , ranging from 0.42 to 0.51 despite comparable ICC_{2k} values. Arguably, reflecting good relative consistency

but weaker absolute agreement. Correlational analyses further revealed that human-AI agreement is strongest for the total score ($P_r = 0.654$) but notably lower for the Coherence and Cohesion and Lexical Resource criteria ($P_r = 0.513$ and $S_\rho = 0.553$ respectively). Thus, pedagogical applications should position AI as a complementary diagnostic tool—particularly for lower-order features—while entrusting nuanced evaluative judgments (e.g., rhetorical effectiveness, creativity) to human raters.

5.1.2.3. Adoption trajectory and future outlook

A robust optimism pervades both literature and the instructors survey: 94.80% anticipate AI becoming a common assessment tool within 5 to 10 years, and 91.70% assert that human oversight will remain crucial. This aligns with digital transformation narratives emphasizing personalized instruction, data-driven decision-making, and the instructor's evolving role from grader to facilitator (Pedro et al., 2019; Zawacki-Richter et al., 2019). However, scaling adoption responsibly will require transparent model explainability, bias mitigation strategies, and sustained professional development for educators.

5.1.2.4. Implications for practice

To ensure the responsible and effective integration of AI into educational assessment, several key practices should be considered (akin to the announcement of American Federation of Teachers, 2025). First, calibration protocols must be developed through systematic norming sessions that align AI scores with those of human raters. This process is essential for mitigating mean biases in both holistic and analytic scoring systems. Second, targeted integration of AI is recommended, wherein AI tools are primarily used to provide formative feedback on discrete aspects such as grammar, syntax, and organization. In contrast, summative assessments should be reserved to human raters to preserve construct validity and uphold pedagogical integrity. Third, the establishment of ethical frameworks is crucial, particularly in the form of regular bias audits and fairness checks, to protect the interests of linguistically and culturally diverse learners. Lastly, instructor training should be prioritized through workshops focusing on prompt engineering and the interpretability of AI feedback. Such training empowers educators

to act as informed “AI coaches,” capable of guiding students through AI-mediated learning environments.

5.1.3. Divergences in perceived versus actual effectiveness

The findings reveal a pronounced gap between university instructors’ confidence in AI assessment tools and these tools’ empirically measured performance.

5.1.3.1. Surface-level features: Grammar and syntax

Instructors rated AI highly for grammar and syntax checking, with 66.70% deeming it *effective* and only 2.10% calling it *ineffective*. By contrast, AI raters achieved substantially lower reliability compared to human raters on the Grammatical Range and Accuracy criterion ($K_\alpha = 0.456$), almost half that of human raters ($K_\alpha = 0.802$). Moreover, AI raters exhibited a systematic underestimation bias, assigning mean scores 0.54 points lower than humans (95% CI $[-0.80, -0.28]$; paired $t(59) = -4.18$, $p < 0.001$), indicating that despite strong instructor endorsement, AI tools currently fall short of human standards in grammatical accuracy.

5.1.3.2. Higher-order skills: Content, organization, and task response

Although 40.60% of instructors viewed AI tools as effective for content and argument analysis—and 47.90% for organization and structure feedback—neutral responses were notably high (38.50% and 40.60%, respectively), reflecting ambivalence about AI tools’ ability to assess higher-order writing skills. Empirical evaluation confirms this uncertainty: AI raters achieved lower reliability on Coherence and Cohesion (CC) ($K_\alpha = 0.468$) and Task Response (TR) ($K_\alpha = 0.421$) compared to human raters ($K_\alpha = 0.714$ and $K_\alpha = 0.860$ respectively). Furthermore, AI under-scored CC by ≈ 0.5 points (95% CI $[-0.69, -0.21]$; paired $t(59) = -3.71$, $p < 0.001$) and TR by ≈ 1 point (95% CI $[-1.15, -0.67]$; paired $t(59) = -7.62$, $p < 0.001$), demonstrating instructors’ overestimation of AI tools’ sensitivity to rhetorical effectiveness and argumentative depth—a critique long articulated in

the AES literature regarding construct validity and reliance on surface proxies (Perelman, 2014; Y. Yang et al., 2002).

5.1.3.3. Overall scoring: Consistency versus accuracy

Instructors remained largely neutral about AI tools' grading effectiveness—57.30% rated them neutral, while only 13.50% deemed them *effective*—a perspective that reflects actual AI performance. Although AI raters demonstrated excellent consistency in total score assessments ($ICC_{2k} = 0.964$), they systematically under-scored essays by an average of ≈ 2.5 (95% CI $[-3.25, -1.68]$) points on the 36-point scale (paired $t(59) = -6.27$, $p < 0.001$; $C_d = -0.81$). Furthermore, the correlation between AI and human total scores was only moderate ($P_r = 0.654$). These findings suggest that instructors' moderate expectations for AI grading are warranted, given its high consistency but notable bias in scoring.

5.1.3.4. Implications for practice and further research

The observed discrepancies between perceived and actual AI effectiveness suggest an imperative for (a) professional development that accurately calibrates instructor expectations and (b) enhanced transparency in AI scoring algorithms. Given that 91.70% of instructors regard human oversight as important or very important, hybrid assessment models may reconcile efficiency gains with the need for accuracy and fairness. Future research should investigate targeted training interventions that acquaint educators with AI performance profiles and explore algorithmic refinements to mitigate underestimation biases, particularly in assessing higher-order writing skills (Xiao et al., 2025).

5.1.4. Contextual factors in the performance of artificial intelligence raters

The substantial variability among AI raters (e.g., Task Response correlations ranging from $P_r = 0.170$ to $P_r = 0.657$) displays the inadequacy of treating *AI assessment* as a monolithic category. The superior performance of *OpenAI o3-mini* across multiple criteria challenges the

prevailing assumption that model size directly correlates with assessment quality (Brown et al., 2020). This divergence may reflect architectural differences in attention mechanisms or training data stratification, supporting Bender et al.'s (2021) contention that model performance is contingent on design choices beyond parameter count.

The demographic context further explains certain divergences from literature. Turkish instructors' higher-than-expected confidence in AI fairness (41.67% agreement) contrasts with Baker and Hawn's (2022) findings in Anglophone contexts, potentially reflecting different institutional technology adoption patterns. Similarly, the strong consensus on AI's future role (94.79% expecting widespread adoption within 5 to 10 years) exceeds Zawacki-Richter et al.'s (2019) global survey results, suggesting region-specific acceleration of technology acceptance.

5.1.5. Ethical and pedagogical implications

Educators' ethical concerns (45.83% expressing elevated worry) validate critical scholarship on algorithmic bias (Baker & Hawn, 2022; Pedro et al., 2019). The observed demographic homogeneity in essays used in the correlational human-AI raters analysis precluded bias analysis but the scoring variability across models raises transparency concerns consistent with Kim and Jo's (2024) "black box" critique. This necessitates the *fairness-aware training* advocated by Bender et al. (2021), particularly for high-stakes contexts.

The pedagogical implications are multifaceted. Strong instructor acceptance of AI for formative feedback (47.92% indicated effective or more) aligns with Cotos' (2014) advocacy for AWE in drafting processes. However, the limited uptake for higher-order feedback suggests AI currently complements rather than replaces human pedagogical roles, supporting Warschauer and Grimes' (2008) expectation of a hybrid assessment system. Crucially, the consensus on human oversight necessity (91.67% agreement) aligns with Chapelle's (2017) emphasis that technology should function as a complementary source of evidence within professional judgment, rather than as a mechanism to replace or automate evaluative decision-making.

5.2. Conclusion and Suggestions

This thesis set out with the dual intention of (1) investigating university instructors' perceptions, practices, and ethical concerns regarding the use of AI tools in assessing writing, and (2) empirically evaluating the alignment between human and AI raters grading argumentative English as a Second Language (ESL) essays using the IELTS-WT2B as the rubric. Employing a concurrent multimethod quantitative research design, the first research strand drew on survey data from 96 English instructors across Turkish universities to illuminate patterns of AI adoption, perceived effectiveness, ethical considerations, and future outlooks. The second research strand systematically investigated scores assigned by 4 human raters and 25 AI tools on 60 argumentative ESL essays, analyzing inter-rater reliability, score distributions, and rank-order correlations. Rather than reiterate every result, this conclusion seeks to interpret the findings, consider broader implications for theory and practice, critically reflect on the research process, acknowledge limitations, and propose directions for future inquiry.

5.2.1. Key findings

The survey revealed that a substantial majority of university instructors (over 80%) have already experimented with AI tools—most commonly ChatGPT and Grammarly—for facets of writing assessment such as grammar and syntax checking, plagiarism detection, and feedback on organization. Instructors expressed the greatest confidence in AI tools' capacity for objective, form-focused tasks (grammar, syntax, plagiarism), while demonstrating caution toward AI's role in marking and content evaluation. Although neutrality was the modal response on many attitudinal items, positive evaluations consistently outnumbered negative ones, especially when AI assessments were aligned with explicit rubrics. Ethical concerns—particularly around data privacy, algorithmic bias, and pedagogical integrity—were acknowledged by nearly half of respondents. Finally, there was widespread optimism that AI will become a commonplace tool in writing assessment within the next 5 to 10 years, coupled with majority's willingness to integrate such technologies regularly into their own practice.

The correlational analysis confirmed the high reliability of human raters, thereby validating human consensus as the benchmark. AI raters showed only moderate inter-model agreement,

yet their aggregated mean scores exhibited stable rank orderings ($ICC_{2k} \approx 0.95$). Consistently, AI raters assigned lower average scores than human raters, with greater score dispersion. Pearson's correlation coefficient (P_r) values between arithmetic mean of human and AI total scores ranged from 0.159 to 0.750, indicating existence of an alignment but also revealing domain-specific gaps: AI raters excelled most in the Lexical Resource criterion yet lagged in the Coherence and Cohesion criterion. Notably, the *OpenAI o3-mini* LLM outperformed others across individual criteria and overall correlation, suggesting that model architecture can significantly influence grading fidelity.

5.2.2. Interpretation and broader implications

The contrast of instructors' guarded yet generally positive attitudes with the empirical grading results showcases a refined reality: AI can reliably support discrete, rule-based aspects of writing assessment but remains less consistent in capturing higher-order, rhetorical dimensions. The moderate correlations in task and content evaluation highlight that AI—when deployed zero-shot with generic prompts—tends to default to formulaic measures of *fulfillment* rather than interpreting argument sophistication or contextual appropriateness. This gap aligns with instructors' expressed concerns about AI tools' inability to account for disciplinary details, cultural contexts, or individual student voice.

Theoretical contributions of this research include substantiating and extending sociocultural perspectives on technology adoption in education. Instructors' conditional trust—contingent on rubric structure and clear criteria—affirms the importance of transparency and pedagogical scaffolding when integrating AI. From a measurement theory standpoint, the findings invite a re-examination of construct validity in AWE: to what extent do AI scores share construct-irrelevant variance with human judgments? The divergent score distributions and reliability coefficients suggest that AI assessments might inadvertently privilege certain linguistic features (e.g., lexical richness) while under-weighting coherence strategies that human raters privilege.

Practically, this thesis offers actionable insights for stakeholders. First, professional development for instructors should focus on strategies to calibrate AI outputs against personal

grading philosophies and established rubrics. Given instructors' higher trust when AI is rubric-aligned, workshop models could train educators in prompt engineering and tool customization to ensure AI feedback maps onto learning objectives. Second, AI tool developers and assessment designers should prioritize the integration of domain-specific training and fine-tuning, especially for coherence and argumentative quality, to close the gap with human raters. Third, institutions must establish clear ethical guidelines—addressing data governance, bias auditing, and equitable access—to alleviate instructors' concerns and foster responsible AI use in assessment.

5.2.3. Directions for future research

Building on the present work, several avenues warrant exploration. Longitudinal studies could track how sustained AI tool use influences instructors' grading practices, confidence, and ethical stances over time. Experimental research involving prompt-engineering interventions and few-shot learning would clarify how tailored AI configurations can bridge gaps in coherence and higher-order evaluation. Broader comparative studies could examine AI grading across multiple writing genres (e.g., narratives, reports) and disciplinary conventions to assess domain sensitivity. Investigating student perspectives on AI evaluation (in similar nature to Er et al., 2025; Smolansky et al., 2023) and its impact on revision practices would illuminate downstream learning effects. Finally, mixed-methods designs incorporating think-aloud protocols during human and AI grading could reveal the cognitive processes underlying scoring decisions, enriching our understanding of construct alignment.

5.2.4. Summary

In sum, this thesis illuminates the multi-faceted terrain of AI integration in writing assessment. Instructors' attitudes—shifting between optimism and guarded skepticism—reflect both recognition of AI tools' efficiency in mechanical tasks and wariness of its limitations in interpreting rhetorical complexity. Empirical comparisons demonstrate that, while AI tools can approximate human scores on aggregate, distinct discrepancies persist, especially in coherence and task fulfillment. These findings promote the imperative of

human-AI complementarity: leveraging AI tools' strengths in form-focused feedback while preserving human judgment for interpretive nuance. As AI continues to mature, thoughtful integration—grounded in pedagogical clarity, ethical safeguards, and ongoing empirical validation—will be essential to realizing its promise as a transformative yet responsible partner in educational assessment.



REFERENCES

- Alharbi, W. (2023). AI in the foreign language classroom: A pedagogical overview of automated writing assistance tools (M. M. Rahman, Ed.). *Education Research International*, 2023, 1–15. <https://doi.org/10.1155/2023/4253331>
- American Federation of Teachers. (2025, July 7). *AFT to Launch National Academy for AI Instruction with Microsoft, OpenAI, Anthropic and United Federation of Teachers*. Retrieved July 10, 2025, from <https://www.aft.org/press-release/aft-launch-national-academy-ai-instruction-microsoft-openai-anthropic-and-united>
- Andrade, H. G. (2000). Using rubrics to promote thinking and learning. *Educational Leadership*, 57(5), 13–19.
- Anthropic. (2025, May 20). *Anthropic - Models Overview*. Anthropic - Models Overview. Retrieved May 20, 2025, from <https://docs.anthropic.com/en/docs/about-claude/models/overview>
- Attali, Y. (2004). Exploring the feedback and revision features of Criterion. *Journal of Second Language Writing*, 14.
- Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. *The Journal of Technology, Learning and Assessment*, 10(3). Retrieved March 26, 2025, from <https://ejournals.bc.edu/index.php/jtla/article/view/1603>
- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning and Assessment*, 4(3). Retrieved March 26, 2025, from <https://ejournals.bc.edu/index.php/jtla/article/view/1650>
- Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32(4), 1052–1092. <https://doi.org/10.1007/s40593-021-00285-9>
- Barrett, A., & Pack, A. (2023). Not quite eye to A.I.: Student and teacher perspectives on the use of generative artificial intelligence in the writing process. *International Journal of Educational Technology in Higher Education*, 20(1), 1–24. <https://doi.org/10.1186/s41239-023-00427-0>

- Bazerman, C., Rogers, P., Stansell, A., Dean, C., Early, J. S., Lunsford, K., & Null, S. (Eds.). (2012, July 8). *International advances in writing research: Cultures, places, measures*. The WAC Clearinghouse; Parlor Press. <https://doi.org/10.37514/PER-B.2012.0452>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models Be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Ben-Simon, A., & Bennett, R. E. (2007). Toward more substantively meaningful automated essay scoring. *The Journal of Technology, Learning and Assessment*, 6(1). Retrieved March 26, 2025, from <https://ejournals.bc.edu/index.php/jtla/article/view/1631>
- Bouziane, K., & Bouziane, A. (2024). AI versus human effectiveness in essay evaluation. *Discover Education*, 3(1), 201. <https://doi.org/10.1007/s44217-024-00320-6>
- Brewer, J., & Hunter, A. (2006). *Foundations of multimethod research*. SAGE Publications, Inc. <https://doi.org/10.4135/9781412984294>
- British Council. (2024). *Who accepts IELTS?*. IELTS Recognition: Accepted Globally | Take IELTS.. Retrieved November 24, 2024, from <https://takeielts.britishcouncil.org/ielts-recognising-organisations/recognise-ielts>
- Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing*. University Press of Colorado.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020, July 22). *Language models are few-shot learners* (4). arXiv: 2005.14165 [cs]. <https://doi.org/10.48550/arXiv.2005.14165>
- Burstein, J. (2012). Automated essay evaluation and scoring. In *The Encyclopedia of Applied Linguistics*. John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781405198431.wbeall1046>
- Burstein, J., & LaFlair, G. T. (2024, November 4). *Where assessment validation and responsible AI meet*. arXiv: 2411.02577 [cs]. <https://doi.org/10.48550/arXiv.2411.02577>

- Burstein, J. C. (2003, January 30). The e-rater® scoring engine: Automated essay scoring with natural language processing. In *Automated Essay Scoring: A Cross-disciplinary Perspective* (0th ed., pp. 113–121). Routledge. <https://doi.org/10.4324/9781410606860>
- Celce-Murcia, M., Brinton, D., & Snow, M. A. (Eds.). (2014). *Teaching english as a second or foreign language* (Fourth edition). National Geographic Learning [u.a.]
- Chapelle, C. A. (2017). Evaluation of technology and language learning. In *The Handbook of Technology and Second Language Teaching and Learning* (pp. 378–392). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781118914069.ch25>
- Chodorow, M., Gamon, M., & Tetreault, J. (2010). The utility of article and preposition error correction systems for english language learners: Feedback and assessment. *Language Testing*, 27(3), 419–436. <https://doi.org/10.1177/0265532210364391>
- Chyung, S. Y. Y., Roberts, K., Swanson, I., & Hankinson, A. (2017). Evidence-Based Survey Design: The Use of a Midpoint on the Likert Scale. *Performance Improvement*, 56(10), 15–23. <https://doi.org/10.1002/pfi.21727>
- Cohen, J. (2013, May 13). *Statistical power analysis for the behavioral sciences* (0th ed.). Routledge. <https://doi.org/10.4324/9780203771587>
- Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed). Wiley.
- Corder, G. W., & Foreman, D. I. (2014). *Nonparametric statistics: A step-by-step approach* (Second edition). Wiley.
- Cotos, E. (2011). Potential of automated writing evaluation feedback. *CALICO Journal*, 28(2), 420–459. <https://doi.org/10.11139/cj.28.2.420-459>
- Cotos, E. (2014). *Genre-based automated writing evaluation for L2 research writing: From design to evaluation and enhancement* (1. publ). Palgrave Macmillan.
- Cotton, F., & Wilson, K. (2011, April 12). *An investigation of examiner rating of coherence and cohesion in the IELTS academic writing task 2*. IELTS: Research Reports. Retrieved November 24, 2024, from <https://ielts.org/researchers/our-research/research-reports/a-n-investigation-of-examiner-rating-of-coherence-and-cohesion-in-the-ielts-academic-writing-task-2>

- Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)*. Common European Framework of Reference for Languages (CEFR).. Retrieved May 20, 2025, from <https://www.coe.int/en/web/common-european-framework-reference-languages>
- Creswell, J. W. (2013). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (Fourth Indian edition). PHI Learning Private Limited.
OCLC: 961503091.
- Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Cumming, G., & McDougall, A. (2000). Mainstreaming AIED into education? *International Journal of Artificial Intelligence in Education*, 11, 197–207.
- Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, 18(1), 7–24. <https://doi.org/10.1016/j.asw.2012.10.002>
- Deller, J. R., & Hansen, J. (2005). 3 - methods, models, and algorithms for modern speech processing. In W.-K. Chen (Ed.), *The Electrical Engineering Handbook* (pp. 861–890). Academic Press. <https://doi.org/10.1016/B978-012170960-0/50063-3>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May 24). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv: 1810.04805 [cs]. <https://doi.org/10.48550/arXiv.1810.04805>
- Diab, R., & Balaa, L. (2011). Developing detailed rubrics for assessing critique writing: Impact on EFL university students' performance and attitudes. *TESOL Journal*, 2(1), 52–72. <https://doi.org/10.5054/tj.2011.244132>
- Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1). Retrieved March 26, 2025, from <https://ejournals.bc.edu/index.php/jtla/article/view/1640>
- Dobrin, S. I. (2023, May 25). *Talking about generative AI: A guide for educators*. Broadview Press.

- du Boulay, B. (2016). Artificial intelligence as an effective classroom assistant. *IEEE Intelligent Systems*, 31(6), 76–81. <https://doi.org/10.1109/MIS.2016.93>
- Edwin Welch, R., & Frick, T. W. (1993). Computerized adaptive testing in instructional settings. *Educational Technology Research and Development*, 41(3), 47–62. <https://doi.org/10.1007/BF02297357>
- Elliot, N. (2016). A theory of ethics for writing assessment. *Journal of Writing Assessment*, 9(1). Retrieved May 30, 2025, from <https://escholarship.org/uc/item/36t565mm>
- Elliot, S. (2003, January 30). IntelliMetric™: From here to validity. In *Automated Essay Scoring: A Cross-disciplinary Perspective* (0th ed.). Routledge. <https://doi.org/10.4324/9781410606860>
- Er, E., Akçapınar, G., Bayazıt, A., Noroozi, O., & Banihashem, S. K. (2025). Assessing student perceptions and use of instructor versus AI-generated feedback. *British Journal of Educational Technology*, 56(3), 1074–1091. <https://doi.org/10.1111/bjet.13558>
- Ericsson, P. F., & Haswell, R. (Eds.). (2006). *Machine scoring of student essays: Truth and consequences*. All USU Press Publications.
- Etikan, I. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics*, 5(1), 1. <https://doi.org/10.11648/j.ajtas.20160501.11>
- Field, A. (2024). *Discovering statistics using IBM SPSS statistics* (Sixth). Sage Publishing.
- Flodén, J. (2024). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal*, n/a(n/a), 1–24. <https://doi.org/10.1002/berj.4069>
- Foltz, P. W., & Rosenstein, M. (2015). Analysis of a large-scale formative writing assessment system with automated feedback. *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, 339–342. <https://doi.org/10.1145/2724660.2728688>
- Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the intelligent essay assessor. In *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (pp. 68–88). Routledge/Taylor & Francis Group.
- Fowler, F. (2008). *Survey research methods* (4th ed.). SAGE Publications, Inc. <https://doi.org/10.4135/9781452230184>

- Gardner, J., O’Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: ‘breakthrough? Or buncombe and ballyhoo?’ *Journal of Computer Assisted Learning*, 37(5), 1207–1216. <https://doi.org/10.1111/jcal.12577>
- Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology*, 26(2), 5–24. <https://doi.org/10.125/73474>
- Google. (2025, May 20). *Gemini 2.5 Pro Preview - Google Cloud*. Gemini 2.5 Pro Preview - Google Cloud. Retrieved May 20, 2025, from <https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/gemini-2.5-pro-preview-05-06>
- Grimes, D., & 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). Retrieved May 30, 2025, from <https://ejournals.bc.edu/index.php/jtla/article/view/1625>
- Hardy, M. (2021, December 22). *Toward educator-focused automated scoring systems for reading and writing*. arXiv: 2112.11973 [cs]. <https://doi.org/10.48550/arXiv.2112.11973>
- Hawkey, R. (2006). *Impact theory and practice: Studies of the IELTS test and progetto lingue 2000*. Cambridge University Press.
- Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1), 77–89. <https://doi.org/10.1080/19312450709336664>
- Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, 13(1), 18617. <https://doi.org/10.1038/s41598-023-45644-9>
- Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. The Center for Curriculum Redesign. OCLC: 1096296369.
- Holstein, K., Vaughan, J. W., Iii, H. D., Dudík, M., & Wallach, H. (2019, January 7). *Improving fairness in machine learning systems: What do industry practitioners need?* arXiv: 1812.05239[cs]. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. <https://doi.org/10.1145/3290605.3300830>

- Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, 5, e208. <https://doi.org/10.7717/peerj-cs.208>
- Hyland, K. (1990). A genre description of the argumentative essay. *RELJ Journal*, 21(1), 66–78. <https://doi.org/10.1177/003368829002100105>
- Hyland, K. (2019). *Second language writing* (Second edition). Cambridge University Press.
- IELTS. (2023, May). *IELTS writing task 2 band descriptors*. IELTS Writing Band Descriptors. Retrieved November 24, 2024, from <https://s3.eu-west-2.amazonaws.com/ielts-web-static/production/Guides/ielts-writing-band-descriptors.pdf>
- IELTS. (2024a). *IELTS academic format: Writing*. IELTS | IELTS Academic Format: Writing. Retrieved November 24, 2024, from <https://ielts.org/take-a-test/test-types/ielts-academic-test/ielts-academic-format-writing>
- IELTS. (2024b). *IELTS academic test*. IELTS | IELTS Academic Test. Retrieved November 24, 2024, from <https://ielts.org/take-a-test/test-types/ielts-academic-test>
- IELTS. (2024c). *IELTS scoring in detail*. IELTS | IELTS Scoring in Detail. Retrieved November 24, 2024, from <https://ielts.org/organisations/ielts-for-organisations/understanding-ielts-scoring>
- IELTS. (2024d). *Who accepts IELTS?*. IELTS | Who Accepts IELTS?. Retrieved November 24, 2024, from <https://ielts.org/take-a-test/why-choose-ielts/who-accepts-ielts>
- IELTS. (2025). *Understanding your score*. IELTS | Understanding your score. Retrieved July 7, 2025, from <https://ielts.org/take-a-test/your-results/ielts-scoring-in-detail>
- IELTS. (n.d.). *IELTS writing key assessment criteria*. IELTS writing key assessment criteria | IELTS.. Retrieved November 24, 2024, from <https://s3.eu-west-2.amazonaws.com/ielts-web-static/production/Guides/ielts-writing-key-assessment-criteria.pdf>
- Jayaron, J., & Blessy, J. J. (2024). Educators' academic insights on artificial intelligence: Challenges and opportunities. *Electronic Journal of E-learning*, 59–77. <https://doi.org/10.34190/ejel.21.5.3272>
- Jick, T. D. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly*, 24(4), 602–611. <https://doi.org/10.2307/2392366>
- Johnson, R. B., & Christensen, L. B. (2014). *Educational research: Quantitative, qualitative, and mixed approaches* (5. ed). SAGE.

- Joshi, A., Kale, S., Chandel, S., & Pal, D. (2015). Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, 7(4), 396–403. <https://doi.org/10.9734/BJAST/2015/14975>
- Jurafsky, D., & Martin, J. H. (2024). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models* (3rd). <https://web.stanford.edu/~jurafsky/slp3/>
- Kabra, A., Bhatia, M., Kumar, Y., Li, J. J., & Shah, R. R. (2021, November 14). *Evaluation toolkit for robustness testing of automatic essay scoring systems*. arXiv: 2007.06796 [cs]. <https://doi.org/10.48550/arXiv.2007.06796>
- Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art, 6300–6308. Retrieved March 26, 2025, from <https://www.ijcai.org/proceedings/2019/879>
- Kellogg, R. T., Whiteford, A. P., & Quinlan, T. (2010). Does automated feedback help students learn to write? *Journal of Educational Computing Research*, 42(2), 173–196. <https://doi.org/10.2190/EC.42.2.c>
- Kemal, S., & Liman-Kaban, A. (2024). Comparative analysis of human graders and AI in assessing secondary school EFL journal writing. *Asian Journal of Distance Education*. Retrieved March 31, 2025, from <https://www.asianjde.com/ojs/index.php/AsianJDE/article/view/796>
- Kendall, M. G., Stuart, A., & Ord, J. K. (1973). *The advanced theory of statistics. 2: Inference and relationship*. Griffin.
- Khan, I., Chohan, I., & Malik, S. I. (2024). Leveraging ChatGPT-4 for enhanced education: Personalized problem-solving and consistent learning. *2024 2nd International Conference on Computing and Data Analytics (ICCDA)*, 1–6. <https://doi.org/10.1109/ICCDA64887.2024.10867340>
- Kim, S., & Jo, M. (2024). Is GPT-4 alone sufficient for automated essay scoring?: A comparative judgment approach based on rater cognition. *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, 315–319. <https://doi.org/10.1145/3657604.3664703>

- Kizilcec, R. F., Huber, E., Papanastasiou, E. C., Cram, A., Makridis, C. A., Smolansky, A., Zeivots, S., & Radulescu, C. (2024). Perceived impact of generative AI on assessments: Comparing educator and student perspectives in Australia, cyprus, and the United States. *Computers and Education: Artificial Intelligence*, 7, 100269. <https://doi.org/10.1016/j.caeai.2024.100269>
- Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, 6, 100225. <https://doi.org/10.1016/j.caeai.2024.100225>
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023, January 29). *Large language models are zero-shot reasoners*. arXiv: 2205.11916 [cs]. <https://doi.org/10.48550/arXiv.2205.11916>
- Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Krippendorff, K. (2019). *Content analysis: An introduction to its methodology*. SAGE Publications, Inc. <https://doi.org/10.4135/9781071878781>
- Kulas, J. T., Stachowski, A. A., & Haynes, B. A. (2008). Middle response functioning in likert-responses to personality items. *Journal of Business and Psychology*, 22(3), 251–259. <https://doi.org/10.1007/s10869-008-9064-2>
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2–3), 259–284. <https://doi.org/10.1080/01638539809545028>
- Landauer, T. K., Laham, D., & Foltz, P. W. (2002). Automated scoring and annotation of essays with the intelligent essay assessor™. In *Automated Essay Scoring* (pp. 87–112). Routledge.
- Leah, G., Ethel, O., & M., R. L. (2024). Algorithmically-driven writing and academic integrity: Exploring educators’ practices, perceptions, and policies in AI era. *International Journal for Educational Integrity*, 20(1), 3. <https://doi.org/10.1007/s40979-024-00153-8>
- Levin, K. A. (2006). Study design III: Cross-sectional studies. *Evidence-based Dentistry*, 7(1), 24–25. <https://doi.org/10.1038/sj.ebd.6400375>

- Li, Z., Feng, H.-H., & Saricaoglu, A. (2017). The short-term and long-term effects of AWE feedback on ESL students' development of grammatical accuracy. *CALICO Journal*, 34(3), 355–375. <https://doi.org/10.1558/cj.26382>
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22, 5–55.
- Link, S., Mehrzad, M., & Rahimi, M. (2022). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*, 35(4), 605–634. <https://doi.org/10.1080/09588221.2020.1743323>
- Lukianenko, V., & Kornieva, Z. (2024). Generative AI in student essays: English teachers' perspectives on effective assessment methods. *Xlinguae*, 17(4), 235–250. <https://doi.org/10.18355/xl.2024.17.04.14>
- Lundgren, M. (2024, June 24). *Large language models in student assessment: Comparing ChatGPT and human graders*. arXiv: 2406.16510 [econ]. <https://doi.org/10.48550/arXiv.2406.16510>
- Macdonald, N. H., Frase, L. T., Gingrich, P. S., & Keenan, S. A. (1982). The writer's workbench: Computer aids for text analysis. *Educational Psychologist*, 17(3), 172–179. <https://doi.org/10.1080/00461528209529254>
- Merkviladze, M. (2024). Integration of automated feedback tools in EFL academic writing classes: Teachers' perspective. *Journal of Education in Black Sea Region*, 10(1). <https://doi.org/10.31578/jeps.v10i1.323>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, September 7). *Efficient estimation of word representations in vector space*. arXiv: 1301.3781 [cs]. <https://doi.org/10.48550/arXiv.1301.3781>
- Morin, J.-F., Olsson, C., & Atikcan, E. Ö. (Eds.). (2021). *Research methods in the social sciences: An a-Z of key concepts*. Oxford University Press.
- Mulyono, P. (2023). AI writing correction tools: Teachers and students' perception. *Jurnal Tatsqif*, 21(1), 35–66. <https://doi.org/10.20414/jtq.v21i1.7963>
- Nadler, J. T., Weston, R., & Voyles, E. C. (2015). Stuck in the Middle: The Use and Interpretation of Mid-Points in Items on Questionnaires. *The Journal of General Psychology*, 142(2), 71–89. <https://doi.org/10.1080/00221309.2014.994590>

- Nisar Ahmad Koka, A. S. K. (2023). Exploring the attitudes and perceptions of foreign language lecturers on artificial intelligence-driven writing evaluation and feedback tools for improving undergraduate writing skills. *Journal of Southwest Jiaotong University*, 58(5). <https://doi.org/10.35741/issn.0258-2724.58.5.6>
- OpenAI. (2025, May 20). *O3 - Model - OpenAI API*. o3 - Model - OpenAI API.. Retrieved May 20, 2025, from <https://platform.openai.com>
- Page, E. B. (1966). The imminence of... Grading essays by computer. *The Phi Delta Kappan*, 47(5), 238–243. Retrieved January 9, 2025, from <https://www.jstor.org/stable/20371545>
- Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education*, 14(2), 210–225. <https://doi.org/10.1007/BF01419938>
- Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and Policy in Mental Health and Mental Health Services Research*, 42(5), 533–544. <https://doi.org/10.1007/s10488-013-0528-y>
- Pearson, K. (1895). VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347–352), 240–242. <https://doi.org/10.1098/rspl.1895.0041>
- Pedro, F., Subosa, M., Rivas, A., & Valverde, P. (2019). Artificial intelligence in education: Challenges and opportunities for sustainable development. *Ministerio De Educación*. Retrieved January 5, 2025, from <https://repositorio.minedu.gob.pe/handle/20.500.12799/6533>
- Accepted: 2019-03-25T16:14:40Z.
- Perelman, L. (2012). Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES). In C. Bazerman, P. Rogers, A. Stansell, C. Dean, J. S. Early, K. Lunsford, & S. Null (Eds.), *International Advances in Writing Research: Cultures, Places, Measures*. The WAC Clearinghouse; Parlor Press. <https://doi.org/10.37514/PER-B.2012.0452>
- Perelman, L. (2014). When “the state of the art” is counting words. *Assessing Writing*, 21, 104–111. <https://doi.org/10.1016/j.asw.2014.05.001>

- Pratt, J. W. (1959). Remarks on zeros and ties in the wilcoxon signed rank procedures. *Journal of the American Statistical Association*, 54(287), 655–667. <https://doi.org/10.1080/01621459.1959.10501526>
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI Preprint*.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2023, September 19). *Exploring the limits of transfer learning with a unified text-to-text transformer*. arXiv: 1910.10683 [cs]. <https://doi.org/10.48550/arXiv.1910.10683>
- Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, 55(3), 2495–2527. <https://doi.org/10.1007/s10462-021-10068-2>
- 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. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing*, 52, 100816. <https://doi.org/10.1016/j.jslw.2021.100816>
- Razali, N. M., & Yap, B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, 2(1), 21–33.
- Rodriguez, P. U., Jafari, A., & Ormerod, C. M. (2019, September 18). *Language models and automated essay scoring*. arXiv: 1909.09482 [cs]. <https://doi.org/10.48550/arXiv.1909.09482>
- Roe, J., Perkins, M., & Ruelle, D. (2024, June 22). *Understanding student and academic staff perceptions of AI use in assessment and feedback*. arXiv: 2406.15808 [cs]. <https://doi.org/10.70770/rzzz6y35>
- Roscoe, R. D., & McNamara, D. S. (2013). Writing pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*, 105(4), 1010–1025. <https://doi.org/10.1037/a0032340>

- Rudner, L. M., & Liang, T. (2002). Automated essay scoring using bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2). Retrieved February 21, 2025, from <https://ejournals.bc.edu/index.php/jtla/article/view/1668>
- Russell, S. J., & Norvig, P. (with Chang, M.-w., Devlin, J., Dragan, A., Forsyth, D., Goodfellow, I., Malik, J., Mansinghka, V., Pearl, J., & Wooldridge, M. J.). (2022). *Artificial intelligence: A modern approach* (Fourth edition, global edition). Pearson.
- Saldaña, J. (2021). *The coding manual for qualitative researchers* (4e). SAGE.
- Saricaoglu, A., & Bilki, Z. (2021). Voluntary use of automated writing evaluation by content course students. *Recall*, 33(3), 265–277. <https://doi.org/10.1017/S0958344021000021>
- Seßler, K., Fürstenberg, M., Bühler, B., & Kasneci, E. (2024, November 25). *Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring* (1). arXiv: 2411.16337 [cs]. <https://doi.org/10.48550/arXiv.2411.16337>
- Shabani, E. A., & Panahi, J. (2020). Examining consistency among different rubrics for assessing writing. *Language Testing in Asia*, 10(1), 12. <https://doi.org/10.1186/s40468-020-00111-4>
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4), 591. <https://doi.org/10.2307/2333709>
- Shapiro, S. S., Wilk, M. B., & Chen, H. J. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association*, 63(324), 1343. <https://doi.org/10.2307/2285889>
- Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20, 53–76. <https://doi.org/10.1016/j.asw.2013.04.001>
- Shermis, M. D., & Burstein, J. C. (Eds.). (2003, January 30). *Automated essay scoring: A cross-disciplinary perspective* (0th ed.). Routledge. <https://doi.org/10.4324/9781410606860>
- Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In *Handbook of Automated Essay Evaluation* (Online-Ausg). Routledge. <https://doi.org/10.4324/9780203122761>

- Shermis, M. D., & Wilson, J. (2024). Introduction to automated essay evaluation. In *The Routledge International Handbook of Automated Essay Evaluation*. Routledge.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. <https://doi.org/10.1037/0033-2909.86.2.420>
- Smolansky, A., Cram, A., Radulescu, C., Zeivots, S., Huber, E., & Kizilcec, R. F. (2023). Educator and student perspectives on the impact of generative AI on assessments in higher education. *Proceedings of the Tenth ACM Conference on Learning @ Scale*, 378–382. <https://doi.org/10.1145/3573051.3596191>
- Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15(1), 72–101. <https://doi.org/10.2307/1412159>
- Stahl, M., Biermann, L., Nehring, A., & Wachsmuth, H. (2024, April 24). *Exploring LLM prompting strategies for joint essay scoring and feedback generation*. arXiv: 2404.15845 [cs]. <https://doi.org/10.48550/arXiv.2404.15845>
- Stevenson, M. (2016). A critical interpretative synthesis: The integration of automated writing evaluation into classroom writing instruction. *Computers and Composition*, 42, 1–16. <https://doi.org/10.1016/j.compcom.2016.05.001>
- Stevenson, M., & Phakiti, A. (2019). Automated feedback and second language writing. In F. Hyland & K. Hyland (Eds.), *Feedback in Second Language Writing: Contexts and Issues* (2nd ed., pp. 125–142). Cambridge University Press. <https://doi.org/10.1017/9781108635547.009>
- Student. (1908). The probable error of a mean. *Biometrika*, 6(1), 1. <https://doi.org/10.2307/2331554>
- Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In J. Su, K. Duh, & X. Carreras (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1882–1891). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D16-1193>
- Tavakol, M., & Dennick, R. (2011). Making sense of cronbach's alpha. *Int J Med Educ*, 2, 53–55. <https://doi.org/10.5116/ijme.4dfb.8dfd>
- Turnitin. (2023, April 4). *Turnitin's AI writing detector: Launch and future plans*. Retrieved May 20, 2025, from <https://www.turnitin.com/blog/the-launch-of-turnitins-ai-writing-detector-and-the-road-ahead>

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023, August 2). *Attention is all you need*. arXiv: 1706.03762 [cs]. <https://doi.org/10.48550/arXiv.1706.03762>
- Velasco, E. (2023, November 6). *Small corpus of colombian english as a second language essays (SCoCESLE)*. Mendeley Data. <https://doi.org/10.17632/WFCBFY29WM.1>
- Vojak, C., Kline, S., Cope, B., McCarthey, S., & Kalantzis, M. (2011). New spaces and old places: An analysis of writing assessment software. *Computers and Composition*, 28(2), 97–111. <https://doi.org/10.1016/j.compcom.2011.04.004>
- Warschauer, M. (2010). New tools for teaching writing. *Language Learning & Technology*, 14(1), 3–8. <https://doi.org/http://dx.doi.org/10125/44196>
- Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: an International Journal*, 3(1), 22–36. <https://doi.org/10.1080/15544800701771580>
- Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, T. (2023, October 31). *The affordances and contradictions of AI-generated text for writers of english as a second or foreign language*. 4404380. <https://doi.org/10.2139/ssrn.4404380>
- Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, 10(2), 157–180. <https://doi.org/10.1191/1362168806lr190oa>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023, January 10). *Chain-of-thought prompting elicits reasoning in large language models*. arXiv: 2201.11903 [cs]. <https://doi.org/10.48550/arXiv.2201.11903>
- Wetzler, E. L., Cassidy, K. S., Jones, M. J., Frazier, C. R., Korbut, N. A., Sims, C. M., Bowen, S. S., & Wood, M. (2024). Grading the graders: Comparing generative AI and human assessment in essay evaluation. *Teaching of Psychology*, 7. <https://doi.org/10.1177/00986283241282696>
- Weyand, L., Goff, B., & Newell, G. (2018). The social construction of warranting evidence in two classrooms. *Journal of Literacy Research*, 50(1), 97–122. <https://doi.org/10.1177/1086296X17751173>
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80. <https://doi.org/10.2307/3001968>

- Williamson, D. M., Bejar, I. I., & Mislevy, R. J. (Eds.). (2006). *Automated scoring of complex tasks in computer-based testing*. Erlbaum.
- Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1), 2–13. <https://doi.org/10.1111/j.1745-3992.2011.00223.x>
ERIC Number: EJ959585.
- 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>
- Wilson, M. (2006). *Rethinking rubrics in writing assessment*. Heinemann Portsmouth, NH. OCLC: 61513117.
- Wilson, M. (2018). *Reimagining writing assessment: From scales to stories*. Heinemann. OCLC: 1005195877.
- Wresch, W. (1993). The imminence of grading essays by computer—25 years later. *Computers and Composition*, 10(2), 45–58. [https://doi.org/10.1016/S8755-4615\(05\)80058-1](https://doi.org/10.1016/S8755-4615(05)80058-1)
- Xiao, C., Ma, W., Song, Q., Xu, S. X., Zhang, K., Wang, Y., & Fu, Q. (2025). Human-AI collaborative essay scoring: A dual-process framework with LLMs. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 293–305. <https://doi.org/10.1145/3706468.3706507>
- Yang, K., Raković, M., Li, Y., Guan, Q., Gašević, D., & Chen, G. (2024, January 11). *Unveiling the tapestry of automated essay scoring: A comprehensive investigation of accuracy, fairness, and generalizability*. arXiv: 2401.05655 [cs]. <https://doi.org/10.48550/arXiv.2401.05655>
- Yang, Y., Buckendahl, C. W., Juszkievicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education*, 15(4), 391–412. https://doi.org/10.1207/S15324818AME1504_04

- Zamfirescu-Pereira, J., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21. <https://doi.org/10.1145/3544548.3581388>
- 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), 1–27. <https://doi.org/10.1186/s41239-019-0171-0>
- Zhong, Y., Hao, J., Fauss, M., Li, C., & Wang, Y. (2024, November 13). *Evaluating AI-generated essays with GRE analytical writing assessment*. arXiv: 2410.17439 [cs]. <https://doi.org/10.48550/arXiv.2410.17439>

APPENDIX A

Configuration of large language models

The following sections contain the prompts for both the system (developer) and user roles used across all artificial intelligence (AI) raters (hereinafter large language models (LLMs)). Below, the response format requested from LLMs and model configurations are also given.

1. System prompt

All LLMs were given the same following system prompt. The prompt contains verbatim the entirety of IELTS Writing Task 2 Bands (IELTS-WT2B) (IELTS, 2023).

You are an IELTS Writing Task 2 examiner. Use the following "IELTS
↔ Writing Task 2" bands to evaluate the student's response strictly
↔ and only based on these criteria:

1. Task Response:

1.0. Band 0: Should only be used where a candidate did not attend or
↔ attempt the question in any way, used a language other than
↔ English, or where there is proof that a candidate's answer has
↔ been totally memorized.

1.1. Band 1: Responses of 20 words or fewer are rated at Band 1. The
↔ content is wholly unrelated to the prompt. Any copied rubric must
↔ be discounted.

1.2. Band 2: The content is barely related to the prompt. No
↔ position can be identified. There may be glimpses of one or two
↔ ideas without development.

1.3. Band 3: No part of the prompt is adequately addressed, or the
↔ prompt has been misunderstood. No relevant position can be
↔ identified, and/or there is little direct response to the

↔ question(s). There are few ideas, and these may be irrelevant or
↔ insufficiently developed.

1.4. Band 4: The prompt is tackled in a minimal way, or the answer

↔ is tangential, possibly due to some misunderstanding of the
↔ prompt. The format may be inappropriate. A position is
↔ discernible, but the reader has to read carefully to find it.
↔ Main ideas are difficult to identify and such ideas that are
↔ identifiable may lack relevance, clarity and/or support. Large
↔ parts of the response may be repetitive.

1.5. Band 5: The main parts of the prompt are incompletely

↔ addressed. The format may be inappropriate in places. The writer
↔ expresses a position, but the development is not always clear.
↔ Some main ideas are put forward, but they are limited and are not
↔ sufficiently developed and/or there may be irrelevant detail.
↔ There may be some repetition.

1.6. Band 6: The main parts of the prompt are addressed (though some

↔ may be more fully covered than others). An appropriate format is
↔ used. A position is presented that is directly relevant to the
↔ prompt, although the conclusions drawn may be unclear,
↔ unjustified or repetitive. Main ideas are relevant, but some may
↔ be insufficiently developed or may lack clarity, while some
↔ supporting arguments and evidence may be less relevant or
↔ inadequate.

1.7. Band 7: The main parts of the prompt are appropriately

↔ addressed. A clear and developed position is presented. Main
↔ ideas are extended and supported but there may be a tendency to
↔ over-generalize or there may be a lack of focus and precision in
↔ supporting ideas/material.

1.8. Band 8: The prompt is appropriately and sufficiently addressed.

↔ A clear and well-developed position is presented in response to
↔ the question(s). Ideas are relevant, well-extended and supported.
↔ There may be occasional omissions or lapses in content.

1.9. Band 9: The prompt is appropriately addressed and explored in
↪ depth. A clear and fully developed position is presented which
↪ directly answers the question(s). Ideas are relevant, fully
↪ extended and well supported. Any lapses in content or support are
↪ extremely rare.

2. Coherence and Cohesion:

2.0. Band 0: Should only be used where a candidate did not attend or
↪ attempt the question in any way, used a language other than
↪ English, or where there is proof that a candidate's answer has
↪ been totally memorized.

2.1. Band 1: Responses of 20 words or fewer are rated at Band 1. The
↪ writing fails to communicate any message and appears to be by a
↪ virtual non-writer.

2.2. Band 2: There is little relevant message, or the entire
↪ response may be off-topic. There is little evidence of control of
↪ organizational features.

2.3. Band 3: There is no apparent logical organization. Ideas are
↪ discernible but difficult to relate to each other. There is
↪ minimal use of sequencers or cohesive devices. Those used do not
↪ necessarily indicate a logical relationship between ideas. There
↪ is difficulty in identifying referencing. Any attempts at
↪ paragraphing are unhelpful.

2.4. Band 4: Information and ideas are evident but not arranged
↪ coherently and there is no clear progression within the response.
↪ Relationships between ideas can be unclear and/or inadequately
↪ marked. There is some use of basic cohesive devices, which may be
↪ inaccurate or repetitive. There is inaccurate use or a lack of
↪ substitution or referencing. There may be no paragraphing and/or
↪ no clear main topic within paragraphs.

2.5. Band 5: Organization is evident but is not wholly logical and
↪ there may be a lack of overall progression. Nevertheless, there
↪ is a sense of underlying coherence to the response. The
↪ relationship of ideas can be followed but the sentences are not
↪ fluently linked to each other. There may be limited/overuse of
↪ cohesive devices with some inaccuracy. The writing may be
↪ repetitive due to inadequate and/or inaccurate use of reference
↪ and substitution. Paragraphing may be inadequate or missing.

2.6. Band 6: Information and ideas are generally arranged coherently
↪ and there is a clear overall progression. Cohesive devices are
↪ used to some good effect but cohesion within and/or between
↪ sentences may be faulty or mechanical due to misuse, overuse or
↪ omission. The use of reference and substitution may lack
↪ flexibility or clarity and result in some repetition or error.
↪ Paragraphing may not always be logical and/or the central topic
↪ may not always be clear.

2.7. Band 7: Information and ideas are logically organized, and
↪ there is a clear progression throughout the response. (A few
↪ lapses may occur, but these are minor.) A range of cohesive
↪ devices including reference and substitution is used flexibly but
↪ with some inaccuracies or some over/under use. Paragraphing is
↪ generally used effectively to support overall coherence, and the
↪ sequencing of ideas within a paragraph is generally logical.

2.8. Band 8: The message can be followed with ease. Information and
↪ ideas are logically sequenced, and cohesion is well managed.
↪ Occasional lapses in coherence and cohesion may occur.
↪ Paragraphing is used sufficiently and appropriately.

2.9. Band 9: The message can be followed effortlessly. Cohesion is
↪ used in such a way that it very rarely attracts attention. Any
↪ lapses in coherence or cohesion are minimal. Paragraphing is
↪ skillfully managed.

3. Lexical Resource:

3.0. Band 0: Should only be used where a candidate did not attend or

↪ attempt the question in any way, used a language other than

↪ English, or where there is proof that a candidate's answer has

↪ been totally memorized.

3.1. Band 1: Responses of 20 words or fewer are rated at Band 1. No

↪ resource is apparent, except for a few isolated words.

3.2. Band 2: The resource is extremely limited with few recognizable

↪ strings, apart from memorized phrases. There is no apparent

↪ control of word formation and/or spelling.

3.3. Band 3: The resource is inadequate (which may be due to the

↪ response being significantly under-length). Possible

↪ over-dependence on input material or memorized language. Control

↪ of word choice and/or spelling is very limited, and errors

↪ predominate. These errors may severely impede meaning.

3.4. Band 4: The resource is limited and inadequate for or unrelated

↪ to the task. Vocabulary is basic and may be used repetitively.

↪ There may be inappropriate use of lexical chunks (e.g. memorized

↪ phrases, formulaic language and/or language from the input

↪ material). Inappropriate word choice and/or errors in word

↪ formation and/or in spelling may impede meaning.

3.5. Band 5: The resource is limited but minimally adequate for the

↪ task. Simple vocabulary may be used accurately but the range does

↪ not permit much variation in expression. There may be frequent

↪ lapses in the appropriacy of word choice and a lack of

↪ flexibility is apparent in frequent simplifications and/or

↪ repetitions. Errors in spelling and/or word formation may be

↪ noticeable and may cause some difficulty for the reader.

3.6. Band 6: The resource is generally adequate and appropriate for

↪ the task. The meaning is generally clear in spite of a rather

↪ restricted range or a lack of precision in word choice. If the

↪ writer is a risk-taker, there will be a wider range of vocabulary
↪ used but higher degrees of inaccuracy or inappropriacy. There are
↪ some errors in spelling and/or word formation, but these do not
↪ impede communication.

3.7. Band 7: The resource is sufficient to allow some flexibility

↪ and precision. There is some ability to use less common and/or
↪ idiomatic items. An awareness of style and collocation is
↪ evident, though inappropriacies occur. There are only a few
↪ errors in spelling and/or word formation and they do not detract
↪ from overall clarity.

3.9. Band 8: A wide resource is fluently and flexibly used to convey

↪ precise meanings. There is skillful use of uncommon and/or
↪ idiomatic items when appropriate, despite occasional inaccuracies
↪ in word choice and collocation. Occasional errors in spelling
↪ and/or word formation may occur, but have minimal impact on
↪ communication.

3.8. Band 9: Full flexibility and precise use are widely evident. A

↪ wide range of vocabulary is used accurately and appropriately
↪ with very natural and sophisticated control of lexical features.
↪ Minor errors in spelling and word formation are extremely rare
↪ and have minimal impact on communication.

4. Grammatical Range and Accuracy:

4.0. Band 0: Should only be used where a candidate did not attend or

↪ attempt the question in any way, used a language other than
↪ English, or where there is proof that a candidate's answer has
↪ been totally memorized.

4.1. Band 1: Responses of 20 words or fewer are rated at Band 1. No

↪ rateable language is evident.

4.2. Band 2: There is little or no evidence of sentence forms

↪ (except in memorized phrases).

4.3. Band 3: Sentence forms are attempted, but errors in grammar and
↪ punctuation predominate (except in memorized phrases or those
↪ taken from the input material). This prevents most meaning from
↪ coming through. Length may be insufficient to provide evidence of
↪ control of sentence forms.

4.4. Band 4: A very limited range of structures is used. Subordinate
↪ clauses are rare and simple sentences predominate. Some
↪ structures are produced accurately but grammatical errors are
↪ frequent and may impede meaning. Punctuation is often faulty or
↪ inadequate.

4.5. Band 5: The range of structures is limited and rather
↪ repetitive. Although complex sentences are attempted, they tend
↪ to be faulty, and the greatest accuracy is achieved on simple
↪ sentences. Grammatical errors may be frequent and cause some
↪ difficulty for the reader. Punctuation may be faulty.

4.6. Band 6: A mix of simple and complex sentence forms is used but
↪ flexibility is limited. Examples of more complex structures are
↪ not marked by the same level of accuracy as in simple structures.
↪ Errors in grammar and punctuation occur, but rarely impede
↪ communication.

4.7. Band 7: A variety of complex structures is used with some
↪ flexibility and accuracy. Grammar and punctuation are generally
↪ well-controlled, and error-free sentences are frequent. A few
↪ errors in grammar may persist, but these do not impede
↪ communication.

4.8. Band 8: A wide range of structures is flexibly and accurately
↪ used. The majority of sentences are error-free, and punctuation
↪ is well-managed. Occasional, non-systematic errors and
↪ inappropriacies occur, but have minimal impact on communication.

4.9. Band 9: A wide range of structures is used with full
↪ flexibility and control. Punctuation and grammar are used
↪ appropriately throughout. Minor errors are extremely rare and

↪ have minimal impact on communication.

Scores range from 0-9 for each criterion. Total score is the sum of

↪ all criteria (0-36).

Provide scores based on the following input format: "Essay Question:

↪ <Essay question goes here>. Student Response: <Student's response

↪ goes here>".

Answer in JSON using the schema provided.

2. User prompt

The user prompt was constructed as such, using Python 3:

```
1  # Load the essay question and student response from files
2  input_dir = Path("./Essays/")
3
4  essay_question_file = input_dir /
   ↪  f"EssayQuestion_{essay_index}.txt"
5  essay_response_file = input_dir /
   ↪  f"StudentResponse_{essay_index}.txt"
6
7  with open(essay_question_file, "r", encoding="utf-8") as file:
8      essay_question = file.read().strip()
9  with open(essay_response_file, "r", encoding="utf-8") as file:
10     student_response = file.read().strip()
11
12  # Construct the user prompt
13  user_prompt = f"Essay Question: {essay_question}\nStudent
   ↪  Response: {student_response}"
```

3. Response format

When an LLM, was prompted it was required to use structured output and answer using the following class:

```
1 from pydantic import BaseModel, Field
2 from typing import Annotated
3
4 class IELTSGradingResult(BaseModel):
5     task_response: Annotated[int, Field(strict=True, ge=0, le=9,
6         ↳ description="Score for Task Response (0-9)")]
7     coherence_cohesion: Annotated[int, Field(strict=True, ge=0,
8         ↳ le=9, description="Score for Coherence and Cohesion
9         ↳ (0-9)")]
10    lexical_resource: Annotated[int, Field(strict=True, ge=0,
11        ↳ le=9, description="Score for Lexical Resource (0-9)")]
12    grammatical_range_accuracy: Annotated[int,
13        ↳ Field(strict=True, ge=0, le=9, description="Score for
14        ↳ Grammatical Range and Accuracy (0-9)")]
15    total_score: Annotated[int, Field(strict=True, ge=0, le=36,
16        ↳ description="Total score (sum of all criteria, 0-36)")]
17
```

Please note, however, that DeepSeek and Claude 3.7 Sonnet (Thinking Enabled) do not support structured output.

For DeepSeek, the following text was appended to the system prompt:

```
{
  "task_response": {"description": "Score for Task Response (0-9)",
    ↳ "type": "integer", "minimum": 0, "maximum": 9},
  "coherence_cohesion": {"description": "Score for Coherence and
    ↳ Cohesion (0-9)", "type": "integer", "minimum": 0, "maximum": 9},
  "lexical_resource": {"description": "Score for Lexical Resource
    ↳ (0-9)", "type": "integer", "minimum": 0, "maximum": 9},
}
```

```

    "grammatical_range_accuracy": {"description": "Score for
    ↳ Grammatical Range and Accuracy (0-9)", "type": "integer",
    ↳ "minimum": 0, "maximum": 9},
    "total_score": {"description": "Total score (sum of all criteria,
    ↳ 0-36)", "type": "integer", "minimum": 0, "maximum": 36},
  }

```

EXAMPLE JSON OUTPUT:

```

{
  "task_response": 1,
  "coherence_cohesion": 2,
  "lexical_resource": 3,
  "grammatical_range_accuracy": 4,
  "total_score": 10
}

```

For Claude 3.7 Sonnet, a tool definition was given for it to extract the information by itself:

```

1  tools=[
2    {
3      "name": "ielts_writing_task_2_grading",
4      "description": "Returns the IELTS Writing Task 2 grading
      ↳ scores for a given essay.",
5      "input_schema": {
6        "type": "object",
7        "properties": {
8          "task_response": {"description": "Score for Task
          ↳ Response (0-9)", "type": "integer", "minimum": 0,
          ↳ "maximum": 9},
9          "coherence_cohesion": {"description": "Score for
          ↳ Coherence and Cohesion (0-9)", "type": "integer",
          ↳ "minimum": 0, "maximum": 9},

```

```

10     "lexical_resource": {"description": "Score for Lexical
    ↪ Resource (0-9)", "type": "integer", "minimum": 0,
    ↪ "maximum": 9},
11     "grammatical_range_accuracy": {"description": "Score
    ↪ for Grammatical Range and Accuracy (0-9)", "type":
    ↪ "integer", "minimum": 0, "maximum": 9},
12     "total_score": {"description": "Total score (sum of
    ↪ all criteria, 0-36)", "type": "integer",
    ↪ "minimum": 0, "maximum": 36}
13 },
14     "required": ["task_response", "coherence_cohesion",
    ↪ "lexical_resource", "grammatical_range_accuracy",
    ↪ "total_score"]
15 },
16     "cache_control": {"type": "ephemeral"}
17 }
18 ]

```