

T.R.
BOLU ABANT İZZET BAYSAL UNIVERSITY
INSTITUTE OF GRADUATE STUDIES
DEPARTMENT OF FOREIGN LANGUAGE EDUCATION



**A COMPARATIVE ANALYSIS OF AI-TOOLS AND ENGLISH
TEACHERS IN WRITING ASSESSMENT**

MASTER OF ARTS, ENGLISH LANGUAGE TEACHING

BERKAY İSMAİL NACAĞ

THESIS SUPERVISOR
Assoc. Prof. Dr. Sedat AKAYOĞLU

BOLU, 12/27/2024

ACCEPTANCE AND APPROVAL PAGE

The thesis entitled “**A COMPARATIVE ANALYSIS OF AI-TOOLS AND ENGLISH TEACHERS IN WRITING ASSESSMENT**” prepared and submitted by **Berkay İsmail NACAĞ** in partial fulfillment of the requirements for the degree of **Master of Arts** in Department of **FOREIGN LANGUAGE EDUCATION** with unanimity vote is hereby approved. 12/27/2024

Jury Members

Signature

Supervisor

Assoc. Prof. Dr. Sedat Akayoğlu
Bolu Abant İzzet Baysal University

.....

Member

Assoc. Prof. Dr. Anıl Rakıcıoğlu Söylemez
Bolu Abant İzzet Baysal University

.....

Member

Assoc. Prod. Dr. Aysel Sarıcaoğlu Aygan
Social Sciences University of Ankara

.....

Institute of Graduate Studies Approval

Prof. Dr. İbrahim KÜRTÜL

Director of Institute of Graduate Studies

ETHICAL DECLARATION

In this thesis study, which I have prepared in accordance with Bolu Abant İzzet Baysal University, Institute of Graduate Studies' Thesis Writing Guide and Template;

- I have obtained the data, information and documents presented in the thesis within the framework of academic and ethical rules,
- I have presented all information, documents, evaluations, and results in accordance with the rules of scientific ethics and morality,
- I have cited and referenced all the works I have benefited from in the thesis study,
- I have not made any changes to the data used,
- I declare that the work I have presented in this thesis is original, and in any other case, I declare that I accept all loss of rights that may arise against me.

According to the similarity report obtained by applying the filters determined by the directorate of the institute in the Turnitin program related to the thesis, the similarity index of the thesis does not exceed 30%.

Ethical permission was obtained for this study from Human Research Ethics Committee with the number 2023/470

.....
BERKAY İSMAİL NACAĞ

ABSTRACT

A COMPARATIVE ANALYSIS OF AI-TOOLS AND ENGLISH TEACHERS IN WRITING ASSESSMENT

MA THESIS

BERKAY İSMAİL NACAĞ

BOLU ABANT İZZET BAYSAL UNIVERSITY

INSTITUTE OF GRADUATE STUDIES

DEPARTMENT OF FOREIGN LANGUAGE EDUCATION

(SUPERVISOR: ASSOC. PROF. DR. SEDAT AKAYOĞLU)

BOLU, 27 DECEMBER 2024

XII+150 pages

This research focuses on comparing the evaluation of writing tasks by artificial intelligence technologies and English language teachers. The increasing role of advancing technology in all areas of education and the recent surge in the use of AI tools have been the primary motivations for this study. The research aims to thoroughly examine the similarities and differences between the evaluations conducted by AI and those by English language teachers, providing insights into the possible reasons behind these outcomes. Additionally, this research also examined the differences between the AI enhanced chatbots that are ChatGPT, BingAI, and Gemini, in terms of grading and giving feedback. The study uses a mixed-method research design considering grading as quantitative and feedback as qualitative aspect of the data. During the data collection process, ten first-year teacher candidates from Bolu Abant İzzet Baysal University were asked to write three different types of essays which are narrative essays, expository essays, and argumentative essays. These essays were then evaluated by five English language teachers and three AI tools that are ChatGPT, BingAI, and Gemini. This process of evaluation will be conducted using TOEFL IBT Writing evaluation rubric considering four main criteria which are Task Response, Organization, Coherence, and Language Use. In addition to the four criteria, to get more elaborate and comprehensive grading and feedback “Overall” part will be added as another criterion. A comparative analysis of the results was conducted, and the underlying causes are identified. The results indicate that in terms of grading and giving quality feedback, chatbots performed on-par with teachers despite some noticeable differences. The findings of this research offer valuable insights into the advantages and limitations of AI technologies. Furthermore, it will provide a meaningful perspective on how these technologies can be effectively integrated into educational programs.

KEYWORDS: Artificial Intelligence, Chatbots, Writing evaluation, Essay, Automated Writing Evaluation, Feedback, Grading

ÖZET

**İNGİLİZCE YAZILI METİNLERİN YAPAY ZEKÂ VE İNGİLİZCE
ÖĞRETMENLERİ TARAFINDAN DEĞERLENDİRİLMESİNİN
KARŞILAŞTIRMALI ANALİZİ
YÜKSEK LİSANS TEZİ
BERKAY ISMAIL NACAK
BOLU ABANT İZZET BAYSAL ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
YABANCI DİLLER EĞİTİMİ ANA BİLİM DALI
(TEZ DANIŞMANI: DOÇ. DR. SEDAT AKAYOĞLU)
BOLU, 27 ARALIK 2024
XII + 150 sayfa**

Bu araştırma, yazma görevlerinin yapay zekâ teknolojileri ile İngilizce öğretmenleri tarafından değerlendirilmesini karşılaştırmaya odaklanmaktadır. Gelişen teknolojinin eğitimin tüm alanlarındaki artan rolü ve yapay zekâ araçlarının kullanımındaki son artış, bu çalışmanın temel motivasyon kaynakları olmuştur. Araştırma, yapay zekâ tarafından yapılan değerlendirmeler ile İngilizce öğretmenleri tarafından yapılan değerlendirmeler arasındaki benzerlik ve farklılıkları ayrıntılı bir şekilde incelemeyi ve bu sonuçların olası nedenlerine dair içgörüler sağlamayı amaçlamaktadır. Buna ek olarak bu araştırma, ChatGPT, BingAI, ve Gemini isimli yapay zekâ tabanlı geliştirilmiş modellerin de değerlendirme ve geribildirimleri arasındaki farkları da incelemiştir. Çalışma, verilerin niceliksel yönü olarak puanlamayı ve niteliksel yönü olarak geri bildirim ele alan karma yöntem araştırma tasarımını kullanmaktadır. Veri toplama sürecinde, Bolu Abant İzzet Baysal Üniversitesi'nden on birinci sınıf öğretmen adayı, anlatı (narrative), açıklayıcı (expository) ve tartışmacı (argumentative) olmak üzere üç farklı türde deneme yazıları yazmaya davet edilmiştir. Bu denemeler, beş adet İngilizce öğretmeni ve ChatGPT, BingAI ve Gemini olmak üzere üç yapay zekâ aracı tarafından değerlendirilmiştir. Değerlendirme süreci, TOEFL IBT Yazma değerlendirme rubriği kullanılarak gerçekleştirilmiş ve dört ana kriter dikkate alınmıştır: Göreve Yanıt (Task Response), Organizasyon (Organization), Tutarlılık (Coherence) ve Dil Kullanımı (Language Use). Bu dört kriterin yanı sıra, daha ayrıntılı ve kapsamlı bir puanlama ve geri bildirim elde edebilmek için "Genel" (Overall) kısmı da başka bir kriter olarak eklenmiştir. Sonuçlar karşılaştırmalı bir analizle incelenmiş ve altta yatan nedenler belirlenmiştir. Araştırma sonuçları, yapay zekâ tabanlı sohbet botlarının puanlama ve kaliteli geri bildirim verme açısından öğretmenlerle eşit performans sergilediğini, ancak bazı belirgin farklılıklar olduğunu göstermektedir. Bu araştırmanın bulguları, yapay zeka teknolojilerinin avantajları ve sınırlamaları hakkında değerli bilgiler sunmaktadır. Ayrıca, bu teknolojilerin eğitim programlarına nasıl etkili bir şekilde entegre edilebileceğine dair anlamlı bir bakış açısı sağlayacaktır.

ANAHTAR KELİMELELER: Yapay Zekâ, Chat botları, Yazı değerlendirmesi, Kompozisyon, Otomatikleştirilmiş Yazı Değerlendirmesi, Geribildirim, Puanlama

TABLE OF CONTENTS

	<u>Page</u>
ABSTRACT	iv
ÖZET.....	v
TABLE OF CONTENTS.....	vi
LIST OF FIGURES	ix
LIST OF TABLES	x
LIST OF ABBREVIATIONS AND SYMBOLS	xi
ACKNOWLEDGEMENT.....	xii
1. INTRODUCTION	1
1.1 Background of the Study	2
1.2 Statement of the Problem.....	6
1.3 The Purpose of the Study.....	6
1.4 The Significance of the Study.....	7
1.5 Theoretical Framework of the Study and Related Research	8
1.5.1 Theoretical Framework of the Study	9
1.5.2 L2 Writing and L2 Feedback.....	10
1.5.3 Computer Assisted Language Learning (CALL)	11
1.5.4 Technology in language learning	17
1.5.5 Artificial Intelligence and Language Learning.....	24
1.5.6 Interdisciplinary Nature of Artificial Intelligence.....	27
1.5.7 Artificial Intelligence and Language Learning.....	29
1.5.8 CALL and Writing	37
1.5.9 Artificial Intelligence and Writing	39
1.5.10 CALL and Feedback	41
1.5.11 AI and Feedback	44
2. MATERIALS and METHOD	50
2.1 Research Design	50
2.2 Participants	51
2.2.1 Tools	51
2.3 Data Collection	53
2.3.1 Essay Writing	53
2.3.2 Essay Evaluation	54
2.3.3 Data Collection Procedures	56
2.4 Data Analysis.....	58
2.4.1 Quantitative Data Analysis.....	58
2.4.2 Qualitative Data Analysis.....	58
2.5 The Trustworthiness of the Study.....	59

2.5.1 The Role of the Researcher	59
3. RESULTS.....	60
3.1 Analysis of Human Grading	60
3.1.1 Narrative Essays	60
3.1.2 Expository Essays.....	62
3.1.3 Argumentative Essays	62
3.1.4 Conclusion.....	63
3.2 Analysis of Human Feedback.....	64
3.2.1 Narrative Essays	64
3.2.2 Expository Essays.....	65
3.2.3 Argumentative Essays	66
3.2.4 Conclusion.....	66
3.3 Analysis of the Interviews	67
3.4 Analysis of AI Grading.....	71
3.4.1 Narrative Essays	72
3.4.2 Expository Essays.....	72
3.4.3 Argumentative Essays	73
3.4.4 Conclusion.....	73
3.5 Analysis of AI Feedback	73
3.5.1 Narrative Essays	75
3.5.2 Expository Essays.....	78
3.5.3 Argumentative Essays	82
3.5.4 Conclusion.....	83
3.6 The Comparison Between Human and AI.....	84
3.6.1 The Comparison of Gradings	84
3.6.2 The Comparison of Feedback.....	94
3.7 The Comparison of Chatbots	97
3.8 Conclusion	102
4. DISCUSSION.....	103
4.1 Grading of the Essays	107
4.2 Feedback for the Essays.....	109
4.3 Conclusion	110
5. CONCLUSION.....	111
5.1 Implications of the Study.....	111
5.2 Limitations of the Study	113
5.3 Suggestions for Future Research	115
5.4 Conclusion	116
REFERENCES.....	119
APPENDICES	139
Appendix 1 Example of an Evaluation Sheet for Teachers.....	139
Appendix 2 TOEFL IBT Writing Evaluation Rubric.....	140
Appendix 3 Tables of the Descriptive Analyses	141

Appendix 4 Sample Essays and Feedback	144
---	-----



LIST OF FIGURES

	<u>Page</u>
Figure 1.1 Map of Students' Construction (Drijvers, 2015)	13



LIST OF TABLES

Table 3.1 Normality Test of Narrative Essays	85
Table 3.2 Normality Test of Expository Essays.....	86
Table 3.3 Normality Test of Argumentative Essays	87
Table 3.4 Narrative Essays ANOVA Teachers vs. ChatGPT	88
Table 3.5 Expository Essays ANOVA Teachers vs. ChatGPT.....	88
Table 3.6 Argumentative Essays ANOVA Teachers vs. ChatGPT.....	89
Table 3.7 Narrative Essays Freidman Teachers vs. BingAI.....	90
Table 3.8 Expository Essays Freidman Teachers vs. BingAI	90
Table 3.9 Argumentative Essays Freidman Teachers vs. BingAI.....	90
Table 3.10 Narrative Essays Freidman Teachers vs. Gemini.....	91
Table 3.11 Expository Essays Freidman Teachers vs. Gemini	91
Table 3.12 Argumentative Essays Freidman Teachers vs. Gemini.....	92
Table 3.13 Narrative Essays Freidman ChatGPT vs. BingAI.....	97
Table 3.14 Expository Essays Freidman ChatGPT vs. BingAI	98
Table 3.15 Argumentative Essays ANOVA ChatGPT vs. BingAI.....	98
Table 3.16 Narrative Essays Freidman ChatGPT vs. Gemini.....	98
Table 3.17 Expository Essays Freidman ChatGPT vs. Gemini	99
Table 3.18 Argumentative Essays Freidman ChatGPT vs. Gemini	99
Table 3.19 Narrative Essays Freidman BingAI vs. Gemini	100
Table 3.20 Expository Essays Freidman BingAI vs. Gemini.....	100
Table 3.21 Argumentative Essays Freidman BingAI vs. Gemini	100

LIST OF ABBREVIATIONS AND SYMBOLS

AI	: Artificial Intelligence
AWCF	: Automated Writing Corrective Feedback
AWE	: Automated Writing Evaluation
CALL	: Computer Assisted Language Learning
CLEF	: Cross-Language Evaluation Forum
EFL	: English as a Foreign Language
ELF	: English as a Lingua Franca
ELT	: English Language Teaching
ESL	: English as a Second Language
GPT	: Generative Pre-trained Transformer
ICALL	: Intelligent Computer Assisted Language Learning
ITS	: Intelligent Tutoring System
NLP	: Natural Language Processing
SLA	: Second Language Acquisition
SPSS	: Statistical Package for Social Sciences

ACKNOWLEDGEMENT

First and foremost, I would like to express my deepest gratitude to my thesis advisor, Assoc. Prof. Dr. Sedat Akayoğlu, for their invaluable guidance, support, and encouragement throughout the entire process of this research. Additionally, his assistance and aid far exceed this progress to even Bachelor's years of mine. His expertise and insight have been crucial in shaping this work, and I am immensely grateful for his patience and mentorship.

I would also like to extend my sincere gratitude to the member of my thesis committee, Assoc. Prof. Dr. Anıl Rakıcıoğlu-Söylemez, who was not only a jury member but also one of the greatest inspirations to me as an example academic, and Assoc. Prof. Dr. Aysel Sarıcaoğlu-Aygan, for their support, ideas, and constructive criticism, which have greatly contributed to improving the quality of this work and even further works.

Special gratitude goes to my family for their unconditional support and encouragement. Their continuous help and patience to bear the process with me has been one of the biggest strengths throughout this unforgettable journey. My mother, father, and my sister have always tried to understand me and help me anyway they can without even questioning. Their approach will always be remembered as beautiful moments for me.

Additionally, I would like to thank everyone who has helped me to be here today, tomorrow, and yesterday. This process was a culmination of many years and experiences that will be another wonderful memory that I will always acknowledge and remember with a smile.

Lastly, I am ever grateful to my lovely and dear soulmate for her never-ending support, encouragement, and help. She is without a doubt one of the reasons why I was able to finish this project. Her aid and ideas in the ups and downs of this project acted as a great path for me to navigate through the difficult ways of the academy. Overcoming challenges, and achieving a great end together have made it even more meaningful.

1. INTRODUCTION

English is the most spoken language around the world with almost 1,5 billion speakers which means every one person in five speaks English. This is an unbelievable number and a breakthrough in history. English is a “contact language” between people from different cultures, backgrounds that have even different native languages (Firth, 1996; Seidlhofer, 2005). This contact language is known as a *lingua franca*. As Jenkins et al. (2011) mentioned, at different eras in history; Sanskrit, Greek, Latin, Arabic, Portuguese and other languages had been used as *lingua francas*. However, none of them achieved the number of speakers that English now has for many reasons. Thus, English as a Lingua Franca (ELF) became fairly popular especially after the World War II, with great migrations occurring within and throughout the world specifically in Europe. After the foundation of European Union (EU), it was accepted that the language barrier between EU citizens should be removed, and English was chosen as a medium of communication. Although long before that, during the colonization in the late 16th century English has secured its place as a *lingua franca*, but after the increase in demand in the mid-20th century the need for English has exponentially grown. Due to that, teaching English as a foreign language (EFL) or a second language (ESL) has gained popularity all over the globe.

In addition to that English is not only the language of communication but also the language of science, technology, research and many more. And the nature of not only English but all languages, is being intertwined with the world, constantly evolving and changing. With all things considered, the researchers have been suggesting, creating and implementing new strategies, methods, tools to teach the language as effective and successful as possible (see Brown, 2007; Celce-Murcia, 2001; Richards & Rodgers, 2014; Larsen-Freeman & Anderson, 2011). One way that the researchers have been studying for decades is Computer Assisted Language Learning (CALL). CALL studies are covering an important part of English language and teaching. The main reason for that is everchanging disposition of language and technology which are constantly feeding and affecting each other.

For the aforementioned reasons and additional considerations, this section is dedicated to providing a detailed explanation of the background, purpose, and

significance of the study, addressing the research problem, and defining key terminology.

1.1 Background of the Study

English language teaching has seen as an important “global language” for decades now (Firth, 1996). Throughout the years, teaching English has changed and evolved. Especially in the 20th century, language teaching became an individual profession (Richards & Rodgers, 2014). Starting from the early 1900s methods have started to emerge with the purpose of teaching the language most effectively. During that time methods co-aligned with the needs of the era, trends in teaching and popular Second Language Acquisition (SLA) theories, the Grammar-Translation method (GTM), for instance, was created as a result of perceiving language as a mechanical tool during the time or Communicative Language Teaching (CLT) later appeared as a result of changing needs in learning and learners in addition to recognising language not as a mechanical tool but as a more alive that is mainly used for daily communication (A.P.R Howatt, 1984; Brown, 2007; Hunter & Smith, 2012; Kumaravadivelu, 2012; Richards, 1985; T. Howatt, 1997; Jin & Cortazzi, n.d.). Needless to say, many methods, approaches, and techniques have been suggested. Consequently, different methods required different tools to use. Tools generally reflected the method of teaching however, researchers and teachers tried to find the best tool to teach regardless of method. Coincidentally, computers, which are tools that would change the world, started to show itself during that time. Independently from the improvements in language teaching and learning, this period also brought dramatic improvements in technology. Naturally, technology and language were destined to cross paths at some point. As expected, starting from 60s, CALL began to rise. Computers have been revolutionary tools since their appearance and without a doubt CALL has revolutionized the way to teach and learn a language. As Davies et al. (2013) mentioned CALL has been affected not only by SLA theories and concurrent trends of an era, but also by the developments in computer technology. Everchanging state of the technology allowed researchers to constantly improve CALL field creating new tools and benefiting from new tools. As a result of that, CALL has covered a significant part of language teaching, applied linguistic and language teacher education research. Moreover, due to English being a global language and a lingua franca, its’ state is always affected by

computers simultaneously computers are also affected by English language due to similar reasons. Evidently, this brings a field of research with never-ending opportunities and questions to answer, that is CALL.

As for the history of CALL, it is well documented (Davies et al., 2013). From the introduction of technology into the field of English language and teaching to the unprecedented number of new tools and uses of technology in this field, CALL has been an interesting area of research for researchers. The literature on CALL has been sound, and the evolution of CALL can be followed step-by-step. The origins of this phenomenon can be traced to the latter half of the 20th century, particularly towards the conclusion of the 1960s when computer technologies began to establish their presence within educational contexts. Especially after the initiation of the PLATO project (Bitzer, 1986), studies began to find shapes and frameworks to implement in different fields. Initially, when introduced into language learning and teaching, computer-based activities were limited, and from a modern perspective, they were quite primitive. Although considering the period, it was revolutionary to use computers in language teaching and learning. From a historical perspective, Sanders (2013) and Levy (1996) effectively summarise the 1960s and the mid-nineties periods. During this time, it can be seen that the technology is not ubiquitous yet, and its uses are quite limited. Even the beginning of the millennia does not see a dramatic increase in the use of technology. Bax (2003) predicts that CALL would be invisible and normalised in the near future. He mentioned that, for computers to reach their potential they need to be normal and everywhere. Obviously, history has proven Bax right. Web 2.0 tools, in a way, have reshaped how technology can be utilized both within and outside the classroom. Especially with social networking becoming popular and technology, by all means, getting into every house and the pockets of everyone, CALL field has re-found into something more complex, improved and much more useful.

Currently, CALL is used in almost every aspect of language teaching and learning, from teacher education to improving individual skills while learning a language. This was one of the reasons why CALL is popular. Computers are flexible tools; computer-based tools are even more flexible. With the introduction of every new advancement, researchers in the field have constantly found new ways to implement new tools to best benefit in teaching and learning processes. Another reason behind this popularity is to provide personalized plans and experiences for

each learner and even for teachers. Learners can follow their progresses closely with the pace they see fit. They can create their own learning programs according to their needs and strong skills in a way becoming their own agents. Additionally, teachers can benefit from technology whilst creating plans for their learners. This brings another use of CALL technologies; they effectively decrease the overloaded teachers' burdens. Teachers can utilise computer-based; moreover, AI-based tools to help them in administrative processes; these may include preparing lesson plans, quizzes or providing effective feedback to learners. Moreover, with the improvements in technology everything that was mentioned above becoming even more advanced, better, more efficient, and easier. Latest breakthrough in this regard is without a doubt the introduction of the enhanced AI-technologies.

Artificial Intelligence, or widely known as AI, has been an important technological tool for decades. The matter of machine or intelligence is a discussed topic of multiple fields (Fetzer, 1990). Yet, the benefits of AI are doubtlessly overshadowing the philosophical aspect of the topic. Discussion regarding the AI technologies dates back to 80s and holds its importance to the recent (see Boden, 1996; Fetzer, 1990; Hunt, 2014; Jackson, 2019; Nilsson, 1982). AI technologies benefit from machine learning, deep learning or natural language processing (NLP) and more to develop themselves constantly. Consequently, an almost never-ending process of improvement allows AI tools to better every time someone uses or every time, they "learn". These opportunities AI offers draw to attention of researchers from many fields. From medics to chemistry or physics to education in specific language education, benefit from AI technologies and the tools that was enhanced using AI.

Consequently, CALL studies use AI and AI tools to better enhance the opportunities that CALL presented before. Furthermore, AI in language teaching and learning came to a point that it may divide from CALL to Intelligent CALL (ICALL) (Ali, 2020; Kannan & Munday, 2018). Especially NLP that is used by most AI tools, allows even more personalised and individualised language learning opportunities for learners. Researchers, seeing this opportunity, creates different tools, or finds different purposes of already created tools. One great and late example of it is the recently published AI-enhanced language modelled chatbots. OpenAI has published ChatGPT in November 2022, which is an enhanced language model powered by billions of texts, paragraphs, and written products (OpenAI,

2023). This chatbot is able to engage in human-like conversations and provide meaningful responses to almost any kind of questions. Of course, a super-intelligent ever-learning language model proves to be useful in many instances. Researchers since then have been utilising this bot and the ones that were published after ChatGPT such as Gemini, and BingAI that were also utilised in this study, to implement chatbots into ELT (see Irzawati et al., 2024; Kostka & Toncelli, 2023; Mabuan, 2024).

Ability to generate meaningful responses offers many opportunities to researchers, teachers and even learners. One of the fields that best benefit from generative AI in language teaching and learning is without a doubt writing aspect. One of the main skills in language is writing skills. Writing and writing education can be found in multiple contexts from K-12 to college and university level education and even to workplaces at the “heart of the knowledge economy” (Brandt, 2005, p. 166). Juzwik et al. (2006) indicated that demand for written communication was at its peak. Almost two decades later, this demand has not faded, on the contrary it only increases. Thus, writing skills are more important than it was in the past. Writing as a skill or as an aspect of the language is explained by multiple models though the WWC model by Graham (2018) proposes that writing can be improved with reconceptualization. One of the parts of reconceptualization is given feedback. It can be explained as “information provided by an agent regarding aspects of one’s performance or understanding” (Hattie & Timperley, 2007). Feedback in writing is crucial to improve learners’ skills. Thus, quality formative feedback is important in the process of writing. Whether for writing or writing assessment, CALL and ICALL tools have offered ample opportunities for improving efficiency of writing tasks. Especially for writing evaluation Automated Writing Evaluation (AWE) systems have gained popularity throughout the CALL-ICALL history. Moreover, artificially intelligent systems have increased the effectiveness of AWE modals drastically. Currently, generative language models significantly benefit from what AWE was and what AWE is.

Evidently, generative AI models can be an important tool for researchers, teachers and learners in terms of providing meaningful and complete feedback. The research in this area is fresh and there has not been much research; however, the studies that were published show promise. Therefore, it is important to fill the gaps

in this subject to provide more evidence to strengthen to state of AI technologies in ELT more specifically in giving feedback.

1.2 Statement of the Problem

Feedback in language learning has been a topic of many studies. The effectiveness and importance of feedback have been proven with numerous pieces of evidence. Quality feedback provided by especially teachers and mentors are crucial to enhance students' and learners' writing skills. Additionally, grading is also considered as a critical element for students to both see their level of success and how to carry it out further when combined with feedback. As for the writing, it is an essential language skill especially in the recent years. The demand for any type of writing has increased due to the digitization in the world. This of course, brings different types of writing as well. Writing, which was conventionally done by pen and paper, has been almost totally replaced by digital writing. Hence the skills needed for writing are even more important now.

Consequently, feedback and grading stand as an even more critical aspect in improving writing skills. In this regard, research in CALL provide great opportunities for not only writing skills, but skills to improve language holistically. Thus, CALL studies bring forth great tools such as Grammarly, PaperRater, and additional AWE technologies to provide invaluable feedback to students, learners, and writers. Furthermore, recent developments in computer technologies and AI reveal newer and better tools and opportunities. One tool that is fresh and considered significant is chatbots such as ChatGPT. These chatbots are generative language models that are able to engage in human-like conversations and respond to various questions effectively. Benefiting from them in language learning and teaching, especially whilst grading and providing written feedback to students has enormous chances to be successful. It should be an eminent focal point for researchers. If used and improved correctly, without a doubt, ChatGPT and other chatbots, or newer tools that were enhanced with the power of AI will overcome many challenges including the ones that are yet unknown.

1.3 The Purpose of the Study

Following that, this study aims to enlighten the possible differences between teachers and AI-chatbots in terms of evaluating by grading and giving feedback to three different types of essays with a further goal of recognising the benefits of AI to the field of ELT more specifically grading and giving feedback in ELT. The study

does this by first, comparing different teachers' grading and feedback to find average opinions about the essays, then doing the same thing with the used chatbots. This process targets to separate the grades/feedback as by human teachers and AI. Later, second step is to analyse the data and find similarities and possible differences between the teachers and AI. Consequently, this study provides a detailed analysis of the two and opens new paths in the field of ELT presenting invaluable data regarding the further use of AI technologies in this field. In specific, this research aimed to address following essential questions:

1. How do humans grade and give feedback to different types of essays?
2. How do AI-based chatbots grade and give feedback on different types of essays?
3. What are the differences and/or similarities between human grading and feedback to AI-based grading and feedback?
4. Are there any differences among AI tools – ChatGPT, BingAI, and Gemini- in grading and feedback?

By providing sound answers to these questions this study aims to contribute to the field of ELT and find different and digitized ways whilst giving feedback to learners and their writing tasks. Hence the study will subsidise helpful data to CALL and/or ICALL studies that can further increase the use of technology and AI-enhanced tools for teachers and learners.

1.4 The Significance of the Study

This study focuses on exploring the similarities and possible differences between AI-enhanced chatbots and English language teachers whilst providing feedback and grading different essay types. Moreover, this study also explores the potential differences between different chatbots as well. It is a widely known phenomena that providing feedback to students in the process of writing is essential to help students enhance their writing (Graham et al., 2011; McArthur, 2016; Steiss et al., 2024). However, it is a process that takes significant amounts of time especially when dealing with multiple students (Applebee & Langer, 2011; Graham, 2019; Kiuahara et al., 2009; Steiss et al., 2024).

CALL studies here prove to be effective tools and especially Automated Writing Evaluation (AWE) has been a fundamental topic of research for years (Wilson & Roscoe, 2020; Wilson et al., 2022). AWE systems use NLP and AI to evaluate writing and some studies showed that these systems can positively affect

students considering their engagement, efficacy, or writing quality (Graham et al., 2015; Grimes & Warschauer, 2010; Roscoe et al., 2017; Stevenson & Phakiti, 2014; Wilson & Czik, 2016; Wilson & MacArthur, 2024; Zhai & Ma, 2014). On the other hand, “While some AWE systems are as efficient and reliable as human raters when assigning scores to writing, they are typically less accurate, more generic and verbose, and sometimes confusing to feedback recipients.” (Grimes & Warschauer, 2010; Shermis & Wilson, 2024; Steiss et al., 2024; Wilson & MacArthur, 2024). Nevertheless, technological improvements have brought new types of AI systems, generative language models such as ChatGPT has altered the course of AI. These systems do not require specified human training and more cost effective and even more accessible. These new language models are able to provide effective feedback that is “timely, targeted, adaptive, and useful” which is helpful to students to enhance their writing skills (Biber et al., 2011; Graham et al., 2015; Steiss et al., 2024).

ChatGPT and other AI-enhanced generative tools and chatbots shows great potential to provide effective quality feedback to students and learners. However, chatbots are freshly published tools. ChatGPT was published in November 2023 by OpenAI. After the initial publishment, it has reached millions of users in a span of days. With its success, other companies like Google, Bing, Notion and many more have released their versions of generative language models. Consequently, there is a huge gap in literature regarding the effectiveness of these tools and their similarities, and differences to teachers.

Evidently, whilst there is a considerable amount of data about the effectiveness of AWE tools, AI-enhanced tools concerning writing evaluation, there is little research with the latest artificially intelligent chatbots. Moreover, the literature lacks any significant and detailed research reflecting on the similarities and differences of these tools to human teacher evaluators. Therefore, this study aims to provide invaluable data and evidence with the aim of make an addition to eventually fill that gap.

1.5 Theoretical Framework of the Study and Related Research

English language education and teaching benefit from technology and technological developments since English language is a global phenomenon and currently evolving. This evolution of English language results in a concurrent change in the development and alteration in English language teaching practices

not to fall behind of the everchanging nature of the language and technology. As a consequence of these changes, throughout the history of ELT there has been numerous attempts on using technological advancements for teaching purposes. First, there was Computer Assisted Language Learning (CALL) which altered the course of plenty of teaching practices. The use of computers was only the initial point of using technology for language teaching. Development of more technological tools supplied more methods in this field. After CALL there was Mobile Assisted Language Learning (MALL). We can clearly see that technological developments have been a pivot point and language teaching tools has been parallelly developed following the current trend. Especially in situations where technology can be used most effectively such as second language (L2) writing and L2 feedback, technology and technological developments have attracted spotlights. In this part, a comprehensive examination of the use of technology in ELT and evaluation and assessment will be made also one of the most critical developments of the last century will be covered and explained, that is, artificial intelligence.

1.5.1 Theoretical Framework of the Study

The primary objective of this research is to evaluate the practical application of AI tools, specifically ChatGPT, Bing AI, and Gemini AI, in the context of writing assessment, as compared to feedback provided by English language teachers. This study focuses on the functional and comparative analysis of the feedback these tools generate, aiming to assess their effectiveness, reliability, and pedagogical value. Given the applied nature of this research, the emphasis is on empirical data and measurable outcomes rather than on testing or extending existing theoretical frameworks.

While theoretical perspectives on writing assessment, second language acquisition, and artificial intelligence in education could provide a broader academic context, the scope of this study is intentionally focused on practical implications and real-world usability. By prioritizing direct comparisons and practical insights, this approach seeks to offer actionable recommendations for educators and institutions considering the integration of AI tools into their teaching practices. The absence of a formal theoretical framework allows the study to remain flexible and responsive to the complexities of real-world application, ensuring that the findings are grounded in practice rather than abstract theoretical constructs.

1.5.2 L2 Writing and L2 Feedback

Language educators have considered L2 writing to be one of the most challenging skills to master for language learners. The main considered reason for this complexity is the constant interaction of linguistics, cognitive, and cultural factors that separate L2 writing from L1 writing. Silva (1993) highlights that L2 writers often face difficulties when trying to achieve syntactic complexity, lexical diversity, and cohesion as a result of their limited linguistic resources and processing demands. Additionally, Hinkel (2011) mentions that rhetorical conventions and cultural differences can significantly influence L2 learners' writing approaches, necessitating targeted instructional strategies.

The two primary viewpoints that have dominated research on L2 writing development are process-oriented and product-oriented approaches. The iterative nature of writing, which includes drafting, feedback, and rewriting, is highlighted by process-oriented approaches. According to Hyland (2016), learners can greatly benefit from scaffolded practice and systematic instruction as they gradually improve their linguistic accuracy and rhetorical fluency. Product-oriented techniques, on the other hand, place more emphasis on analysing written texts to pinpoint areas that require development. Furthermore, writing has been viewed as a sociocultural or cognitive activity. The goal of numerous research, including Graham's (2018) Writer(s)-Within-Community (WWC) model, was to integrate the two separate facets of writing into a comprehensive approach to writing and writing evaluation. It is important to note that every viewpoint has made a significant contribution to understanding the development of L2 writing, with growing recognition of the role of feedback in mediating these processes.

Needless to say, feedback is vital for L2 writing proficiency, acting as a bridge between learners' current performance and their desired goals. Thus, written corrective feedback (WCF) has received attention for many years. Studies demonstrated its efficacy in improving grammatical accuracy and linguistic precision (Bitchener & Knoch, 2009). Direct feedback is often more effective for addressing surface-level errors, while indirect feedback can facilitate deeper cognitive engagement and long-term learning. Another term which was considered as a valuable tool is peer feedback. Studies showed that peer feedback encourages collaborative learning and enhances critical thinking skills (Lundstorm & Baker, 2009). However, it is also emphasized that the success of feedback depends on

multiple factors such as learners' feedback literacy, motivation, and ability to integrate feedback into their revisions (Winstone et al., 2017).

As for all things, feedback too has benefited from the advancements of technology. Technological developments have introduced a vast area for writing instruction, providing automated feedback tools as supplementary aids. Automated Writing Evaluation (AWE) tools such as Grammarly and other AI-based systems have revolutionized how learners receive feedback, offering immediate, and detailed suggestions on grammar, vocabulary, and style (Zhang & Hyland, 2018). Literature has found these tools particularly effective especially in terms of addressing simple errors, providing scalable and accessible feedback for learners. However, their limitations are also evident. Ranalli et al. (2022) argue that automated tools often lack the contextual and cultural understanding needed to guide learners on higher-order concerns such as coherence, argumentation, and audience awareness. Furthermore, studies also claimed that while these tools can complement teacher feedback, they are not substitutes for nuanced insights provided by experienced instructors.

All in all, L2 writing and L2 feedback have been considered as critical aspects of language learning and teaching studies revolving around them empathized their importance for the improvement of learners' writing skills. Needless to say, both L2 writing and L2 feedback have benefited from the advancement of technology in increasing the skills of learners by incorporating various tools into the equation. Consequently, Computer-assisted language learning (CALL) has attracted considerable attention in language learning and teaching, including the aspects such as L2 related skills.

1.5.3 Computer Assisted Language Learning (CALL)

In a world where technology is omnipresent and omniscient, one of the oldest and most influential social sciences in history, which is education, must benefit from this ubiquitous tool. In order to do that, researchers have been putting a considerable amount of effort and resources into this field. By the end of 50s a physicist, Chalmers W. Sherwin proposed a computerized learning system at the University of Illinois. After about a year of discussion and clash of ideas, the first version of the PLATO Project started operating on the ILLIAC I computer (Bitzer, 1986). The project quickly became popular and researchers from different fields

tried to incorporate PLATO into their teaching and learning curricula. Hart (1995a) was one of the people who worked on the PLATO IV: Foreign Languages Project. It is an indisputable fact that the project was a milestone in discovering a different side of technology and utilizing it for educational purposes. In the present day, technology has been used for almost half a century to make all parts of education better, easier, and perhaps faster. Many disciplines such as mathematics, science, medicine, engineering, and of course language teaching are benefiting from the vast opportunities of technological tools and devices. Furthermore, there are specific tools and devices that were developed and created for teaching the disciplines and actively using them within the disciplines. To provide a better understanding of the situation, in 1984 only 0.008 percent of the students had access to computers in public schools in the United States (Coley et al., 1997). In recent years it is observed that too many schools have started to use 1:1 computing method (Bebell & O'Dwyer, 2010). Clearly access to computers and technology has been increasing both in public and private schools, and also at homes. Furthermore, computers, phones, tablets, and devices outnumber the number of populations at most homes, and it looks like the number is just going to increase every other day. One of the reasons for this happening is due to the ever-growing pace of technological evolution. Technology is evolving like we are, but at a faster rate. As a result, people always try to keep up with it and adapt it to better use it in daily and professional lives. Allenby and Sarew (2011, p. 35) explains the situation of technology as follows:

Technology is not just a matter of innovation; it is also a matter of adoption by a critical mass of users, and it co-evolves with cultural, economic, political and other domains, each of which continually affects, and is in turn affected by the other.

Whether technology should be used or included in education is somehow still debated by several parties. However, general consensus is that like all other domains, education too must adopt the technology to respond to the current needs of the age. It is one of the ways in which education and language teaching can further improve to better address the current needs of society.

This process of adoption comes in many shapes and sizes. As aforesaid, different subjects use different tools, apps, and devices. For instance, Boon (2009) studies the prospects for teaching 3D geometry using online applets or in another

example Bakker (Bakker, 2004; Bakker & Gravemeijer, 2006; Bakker & Hoffmann, 2005) examined early statistical reasoning of students. Students explore data from two different batteries using apps to design convenient representations and symbolizations. Even simpler, calculators are an essential part of arithmetic studies. In another study by Wijers et al. (2010), teams of grade 7 and grade 8 students play a game using a self-created mobile app called the MobileMath in which they use geometry to construct geometrical shapes using the GPS function in phones and destroying other groups geometrical shapes. The game aims to make students experience geometrical shape in a different way they would in classrooms. Students walk around outside using maps and imagining how they could create shapes whilst destroying the other teams' geometrical shapes.

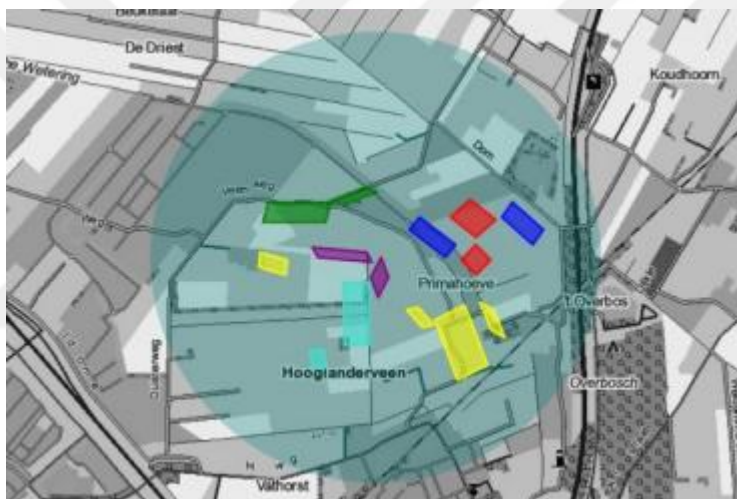


Figure 1.1 Map of Students' Construction (Drijvers, 2015)

There are many such examples of using technology in mathematics education. One of the main aims of this is to create tools that can be reachable whenever and wherever possible, and to help students contextualise the subjects and increase authenticity. Moreover, studies also show that students feel more engaged, and they have more fun using technology even for educational purposes.

One other discipline that benefits much from technology is of course science education. Fundamentally, science is technology and technology is science. Physics, chemistry, biology and all the other natural sciences are quite interactive and basically, they are nature. Seeing, feeling, and experiencing them firsthand is generally more productive than learning only the theories in

classrooms. Furthermore, in a study conducted by Radhamani et al. (2014), even virtual labs are beneficial in enhancing student learning and interaction. Linn (2003), mentioned that:

New technologies, advances in understanding learning, new designs for instruction, and new fields of science jointly push the educational enterprise. Many research programs combine these traditions to improve science learning for teachers, students, and lifelong learners. The past 25 years have yielded a startling assortment of advantageous, disastrous, and intriguing findings. I look forward to the next 25 years in science and technology education.

It has not been 25 years since this paper but the situation of technology in education has drastically changed mostly positively. M.C. Kim et al. (2007) says that “The use of technology, both by teachers to teach and do science and by students to learn and inquire about science, have become core approaches to promoting scientifically and technologically literate citizens...” Research indicate that technology in science education is necessary to provide a better education service to students and to ease teachers’ workload. Consequently, researchers have suggested a variety of technology-enhanced inquiry-oriented approaches (Barab & Luehmann, 2003; Kim & Hannafin, 2004). Courts and Tucker (2012) provides more inclusive data on how to use technology in classrooms. They propose multimedia tools to enhance classroom atmosphere, “There are numerous multimedia tools which can be used to enhance the learning experience in the college classroom.” National Teacher Treaning Institue (NTTI, 2011) indicates that using videos in the classroom increases retention of information by students and helps them to understand the material faster. Following that, Courts and Tucker (2012) put forward some options such as audio, video, and blogging to use in the classrooms.

With time, proposed activities and options have improved by a mile as a recent paper indicates “In contemporary educational settings, the integration of technology has become a pivotal aspect in shaping learning experiences and fostering equitable access to education.” (Eden et al., 2024) In the present context, it is now talked how to integrate emerging technologies like artificial intelligence, augmented reality, virtual reality, virtual worlds, and technologically enhanced tools into classrooms (Rane et al., 2023). From science education to art education, music education to language education technology strengthens its usability every

other day. Although that is the case, why use technology at all? Another question is that accepting the premise of using technology, how to use it effectively.

First of all, even though the robust evidence, data, and research there are still minute concerns on using technology especially some specific occasions and some specific disquiets that arose particularly lately such as data privacy, reliability, or credibility. However, it is generally agreed that as Eden et al. (2024) marked, “Technology has undergone a remarkable evolution within educational settings, revolutionizing traditional teaching methods and significantly impacting the learning landscape.” Technology integration is essential since there are many advantages for learning and teaching. Technology is a great tool to enhance student engagement and motivation because of the rich multimedia opportunities that it offers that captivate the students’ interest (Eden et al., 2024). Also, the immense spread of smartphones, tablets, notebooks, or any mobile device greatly enhances omnipresent learning opportunities thus, resulting in the emergence of formal and informal learning spaces. Furthermore, accessibility of the internet transformed the availability of knowledge and global, cultural interactions, facilitating students to access an almost unlimited amount of information and cooperate with peers and experts around the world (Lock & Redmond, 2021). Within classroom settings, simulations, virtual labs and educational games provide a fun learning platform for students whilst making the lessons more memorable and nurturing a deeper comprehension of concepts for students (Asad et al., 2021). Moreover, the range of different activities support individual learning responding to the various needs of each student. Additionally, with new technologies like artificial intelligence, AI enhanced tools can adapt and evolve according to the student's needs and the interactions between the students and the tool with machine learning algorithms. Also, technology provides educational opportunities and resources especially for students in remote areas (Alkureishi et al., 2021). “Online courses, digital textbooks, and open educational resources democratize access to high-quality education, breaking down barriers of geography, time, and socioeconomic status.” (Eden et al., 2024). In summary, current evidence highlights the importance of the integration of technology in classrooms and lesson plans. For teachers, it is not only crucial not to fall behind with the world around; the more important issue is to catch up with the students, who are born into this technology and native users of it. The evidence also suggests that the more the teachers can relate to the students the more

efficient and teaching becomes. Perhaps if not the most, one of the most important reasons for integrating technology into learning and teaching is to be able to effectively communicate with students in their digitized worlds and minds.

As it was aforementioned, English Language Teaching has always been interconnected with other fields to improve teaching, learning and education in general. After computers were first presented and used as a tool to enhance language learning it quickly became popular to utilise technology and computers to assist in teaching programmes. Therefore, it led to the introduction of one of the most important language learning mediums that is Computer Assisted Language Learning (CALL). First, what is CALL? Computer Assisted Language Learning is defined by Levy (1996) as the study of computer applications in language learning. History of CALL can be traced back to the 1960s to the Illinois University with the aforementioned PLATO project (Programmed Logic for Automated Teaching Operations) (Hart, 1995). The PLATO project was first introduced to integrate computational technologies to the educational field. After the advancement of the computer technologies, many fields including CALL were branched out under the frame of computer technologies in education. Huge appeal in computer technologies led to further improvement of CALL field mostly because of its usefulness. Current literature considers CALL as an umbrella term that is related to the all technological, computing based devices used specifically for language learning and teaching. In their chapter Rakıcıoğlu-Söylemez and Akayoğlu (2019) shows three different focused perspectives of the field of CALL lately as the theoretical foundations and proposed applications, use and effectiveness of CALL practices in the language classrooms, and using CALL as a language teaching tool, lastly the perceptions of the stakeholders. Although this was written in 2016, it is obvious that CALL was a major research area, and since then it is even moving further and further. For instance, Bax (2003) claims that the aim can be the normalisation process that makes CALL finally invisible allowing it to serve the needs of learners and combined into everyday practice for teachers, regarding the future of CALL. Not even twenty years later from this paper, it can be claimed that the future he was thinking of is the present that we are living. Furthermore, it is even bigger than the thought future. Although CALL is a huge part of the teaching and learning languages, it still has immense potential to grow even more. Ima and Jihad (2024) suggest “By harnessing the power of technology, CALL has the

potential to revolutionize language education, fostering inclusive, interactive, and personalized learning experiences in the digital era.” Whilst the studies have improved quite profoundly digitilasing the normal life is rather new. It is only natural for CALL to be further improved in this era of digital age even faster and stronger than before with new technologies being introduced every other day. Before exploring the new and improved, it is crucial to investigate the history of technology in education and language learning in detail.

1.5.4 Technology in language learning

“Language and technology have been deeply intertwined ever since the invention of writing some 5,000 years ago.” (Chun et al., 2016) Languages, just like technology, are a living and breathing phenomenon. As Chun et al. (2016) greatly put above, language and technology feed off each other. This is one of the most important reasons why they are so connected and should be used together. Before diving into the vast relationships of technology and language teaching, what is considered as a use of technology in language studies should be understood. In a paper by Salaberry (2001) it is classified as audiovisual media such as audio, films, videos, and television and teaching tools such as overhead projectors. In more than twenty years, there have been numerous additions to that list. Smartboards, mobile phones, AI enhanced tools, mobile apps, computers, and many more are now used in language teaching and learning. All in all, devices, tools, applications, and anything computer and technological system related can be considered under this name and used as an efficient tool for language learning.

The leading umbrella term that focuses on the relation of technology and language learning is named Computer Assisted Language Learning (CALL) as was mentioned. “As a field that emerged towards the end of the 1970s, CALL has been of interest to the researchers studying language learning and technology.” Akayoğlu (2019) shortly described the beginning of CALL studies as such. After this sudden impact of CALL in the field, Warschauer and Healey (1998) depicted CALL under three phases that are Behavioristic, Communicative, and Integrative. It can be claimed that CALL was following the current and dominant trends in language teaching of the time closely. For instance, Behavioristic CALL was the first phase, then the Communicative CALL as the emergence of Communicative Language Teaching around the globe affected the way of language teaching, (Akayoğlu, 2019) lastly the Integrative CALL with the beginning of the famous post-method era by

Kumaravadivelu (1994). As can be seen and explained by Davies et al. (2013) “Historically computer-assisted language learning (CALL) has been shaped not only by trends in language pedagogy and second language acquisition (SLA) theories, but also by the state of computer technology.” Current literature has a substantial amount of study based around CALL, it covers a significant amount of spot in language studies, but that arouses the question of how did CALL become CALL as it is known at the moment?

The history of the term CALL is rather ambiguous because there is no clear data on when the term first appeared. However, the term was first used by Davies and Steel (1981) in a paper and by 1982 it was a pretty common term in the UK. Later, TESOL started to use the term just a year after Davies and Higgins (1982) published their study (Davies et al., 2013). However, long before CALL was CALL, there was uses of computers for language learning in the 60s. As was discussed above, after the PLATO project gained popularity, language teaching tried to benefit from computer technologies as well. Nonetheless, in their earliest years computers meant to take the workload of teachers by completing repetitive tasks allowing them to focus on the communicative activities (Davies et al., 2013). The PLATO project peaked in the late 70s with PLATO IV. With the programming language named as TUTOR, materials were designed to teach and practice a considerable number of languages such as German, Latin, Russian, ESL, Chinese, and more. These materials included both grammar and communicative related exercises. The University of Illinois increased the number of installed PLATO terminals. There were over 50.000 hours of language practice that had been done by the end of the 70s. (Hart, 1995) Early 70s began with another project by the University of Texas and Brigham Young University called Time-Shared Interactive Computer Controlled Information Television in other words TICCIT Project. (Anderson, 1976; Jones, 1995) This project aimed to develop educative materials for English using computers and televisions. In simple terms, the student would pick a video then an operator system would play the video using televisions. (Anderson, 1976) What distinguished TICCIT from other projects was that the freedom students had using TICCIT, they could move freely going backwards and forwards, skipping some activities whereas other projects had a linear structure. (Jones, 1995) Despite now looking at it, this may feel primitive, but it was an incredible view or activity for the students back then moreover, even the topics that

are considered important now such as autonomous learning, flexible learning was a part of the program. During this period many authoring tools were created similar to TICCIT. CALIS (Hussein et al., 1980), Dasher (Pusack & Otto, 1983) DECU/TUCO (Taylor, 1984) and the Course in Medical and Technical Terminology (Tebben, 1979) were a few crucial projects of them. As can be seen, early CALL was mostly about authoring projects aiming to help teachers create a more authentic, contemporary, and student-based learning environment. Limitations of early technology were also the limitation for CALL tools. With the introduction of microcomputers, CALL tools found a cheaper platform that also supports “enhanced” graphics, animation, and foreign characters. Then the biggest breakthrough of that time came with the internet. Initially, because of older technologies the internet was challenging to utilise, but the first part of the twenty-first century solved those problems with mobile devices and creative tools. (O’Reilly, 2005)

Especially after the microcomputers being affordable, the impact of technology has transitioned into the early stages of education as well. Primary and secondary schools started using technology in classes to provide computerized language learning. The major issue of that time regarding computers and computer programs, was compatibility issues. Computer services and manufacturers had their own operating systems which made the jobs of the programmers working to create language learning tools a lot harder, since the programs could not get exchanged between different operating systems easily (Davies et al., 2013). This ample amount of development of microcomputers also affected the academic circle leading to plenty of CALL publications (Ahmad et al., 1987; Davies & Higgins, 1982; Higgins & Johns, 1984; Hope et al., 1984; Kenning & Kenning, 1984). This period was also the beginning of professional CALL associations. CALICO was founded in 1982 in the United States and EUROCALL in Europe in 1986. With the funding of the European Commission (EU) EUROCALL became a formal professional association. (Davies et al., 2013) Early microcomputers presented extensive text-based practices although they had limited visual opportunities since they were monochrome displays and had inadequate graphic options. Rise of the communicative language teaching approach through that time had little to no effect on CALL productions since there were a richness of the grammar-based activities. Even though that was the case, there were some really creative, imaginative

programs like TUCO (1987) and CLEF (1985) that offered a semi-intelligent approach using tutorial chains, error analysis and feedback. Furthermore, some CALL developers tried to explore artificial intelligence to process students' natural language responses. There were even microcomputer-based AI software, games for communicative practice, *Juegos Comunicativos* a Spanish communicative practice software (Bassein & Underwood, 1985) and *Spion* (Sanders & Sanders, 2013) a German spy game were great examples that greatly resonated with the needs of the communicative approach. With the English being the most dominant language of that time, English as a Foreign Language (EFL) studies saw considerable amount of attention and new methods to teach English were emerging coincidentally matching the opportunities that computer technologies had to offer. For instance, task-based learning (TBL) and digital technologies were great pairs and implementation of technology for methodologies like TBL were a lot easier. Main trend of language teaching during that time was to focus on authenticity in every step of the learning process from contents to tasks or contexts to goals. Flexibility and wide range of uses of digital technologies led to the researchers seeing the potential of technology to solve some practical problems that was faced. They were great tools that could expose learners to 'thinking tasks' with their advanced tools like videos, audiotapes, or flexibility and responsiveness. Traditional activities were replaced with exploratory, real tasks and activities, even when dealing with grammar authentic materials were implemented, learner-centered activities became a priority, and there was a focus on the computer-assisted cognitive tools. (Davies et al., 2013) Language learning process that was based on task-oriented materials with CALL services became the customary practice from then on and later examples could be considered as following the footsteps of this initial process. (Rüschhoff, 2002a, 2013) Authenticity in language learning described by Little (1989, p.5) as creating opportunities for learners to psychologically interact with the target language, "by which we mean the psychological processing of target language input in such a way that it interlocks with modifies the learner's existing knowledge." Also, during this time, discovery-based activities were one other main focus of CALL technologies, furthermore this era had other developments in word-processors, desktop publishing software, databases, information and communication technologies that eventually led those generic applications to be benefited in CALL fields by researchers and teachers as well. One problem during

that time was that the computers did not have recorders or playback technology. In order to record or playback a sound additional tools were needed to be used.

The 1980s were a period where one of the highest profile CALL projects was launched, which were videodisc simulations that aspired to deliver immersion and exposure in the target language. *Montevidisco*, for Spanish learners by BYU (Schneider & Bennion, 1984) and *A la rencontre de Philippe*, for French learners by MIT were the biggest of all videodisc projects. A survey by Rubin et al. (1990) showed that there were seventy-two interactive programs that were videodisc-based. Additionally, *The Domesday* videodisc that was published by BBC in 1986 despite not being published as a language teaching tool recognised as an important and invaluable resource for EFL teachers. With the emergence of new technologies such as CD-ROMs, DVDs in 1990s, and eventually streaming services later on, also the failure of videodiscs in commercial markets closed the period of videodiscs at the beginning of the 90s. This decade redefined the term CALL as Levy (1996) proposed “the search for and study of applications of the computer in language teaching and learning” explanation for CALL, even now this explanation is referred and used in CALL studies. As aforementioned three phases of CALL were identified by Warschauer and Healey (1998) as:

Behaviourist CALL: This phase was mostly implemented from 60s to the end of 70s. The computer was the tutor. In this type of CALL the computer played a role to deliver materials to students. Focus was on drill-and-practice programs.

Communicative CALL: 70s was the beginning of communicative CALL and in the 80s it was the dominant phase. In this phase, computer was again a tool to practice language skills, however this format obtained a format that gave a greater chance of student choice, control, and interaction.

Integrative CALL: Beginning of multimedia and internet defined this phase and became well-known by the mid-90s.

Nonetheless, Bax (2003) had other explanations for CALL. Three approaches of CALL such are Restricted CALL, Open CALL, and Integrated CALL was offered by Bax to consent much more detailed analysis and critical exploration than all other explanations.

Regardless of the explanations, this 90s period was a major threshold for the further development of CALL studies. Multimedia PC arrived that could record and

play sounds that was waited for a long time by teachers and researchers. With these new opportunities CALL practices got further away from drill-and-practice programs. There was talking books which presented text and sound in three different languages. The students could switch between languages and interact with on-screen objects that triggers animations and sound effects. One of the reasons of this breakthrough was the introduction of CD-ROMS that started low on video quality but caught up eventually but offered a higher quality of audio, furthermore it allowed students to engage in activities. Of course, it is necessary to mention that these interactions were not similar to the ones that can be done at present, it was a lot more simple and perhaps primitive from current perspective.

But the most important and life-changing breakthrough was yet to come which was the World Wide Web. Tim Berners-Lee first developed the system to be a closed system at first whilst working at the CERN. It was with the first web browser *Mosaic* (1993) then followed by *Netscape* (1994) that made web public. Initially web was a tool to find resources. Websites were not vivid as they are now, most of them just had texts, some had texts and images. (Davies et al., 2013) Now it is called as Web 1.0 because there was only one way interaction, web users could not directly change or affect the content, they could only view, and they had limited responses to their actions. Consequently, there was limited interaction on the web at the beginning. Virtual learning environments such as *Blackboard* became quite popular during this decade. With the internet developing and users gaining access to a lot faster and more efficient connections, occasions that offered beyond basic grammar exercises was accessible for teachers and learners. MUDs and MOOs that are multi-user-domains, multi-user-domains object oriented, were designed. They were text-based, role-playing adventure games and teachers took this chance to use them as a tool for collaboration and education. Von der Emde et al. (2001) and Shield (2003) explained that how MOOs were used and still being used as language learning instruments.

Beginning of 2000s brought more improvements on computer technologies from video and audio quality to functions of web and access to web. Universities, governments, and even companies had offered online language learning courses. For instance, BBC(www.bbc.co.uk/languages), Chinese online by East China Normal University(www.hanyu.com.cn), LangNET by the US Departments of Defense, and the CAMILLE Group's Ingenio Project was a few of them that used

internet as a medium of language teaching platform. Here the e-learning was redefined as a supporter of traditional learning, whereas it was thought that it could replace the traditional learning beforehand. (Davies et al., 2013) In 2004 Web 2.0 was introduced as a social platform for collaboration, knowledge sharing and networking (O'Reilly, 2005) Web 2.0 provided two-way interaction amongst users. They could comment, share ideas, and interact with each other both synchronously and asynchronously. Just as it was introduced there was a huge interest towards web tools and a tremendous increase in the quantity of online communities that utilised Web 2.0 tools such as blogs, wikis, podcast, social networking websites, virtual worlds, discussion lists, and MUVES that promote sharing, collaboration and interaction (Thomas, 2009). Web 2.0 has opened newer ways of learning a language as digital media gained popularity and took its place in daily lives. Social media platforms allowed people from around the world to interact with each other increasing exposure and authenticity of the targeted language. Moreover, it created fresh opportunities in research in CALL fields as well. For instance, Second Life a 3D virtual world that was launched in 2003 attracted language teachers. In 2005, the first language school was opened in Second Life known as LanguageLab. Cooke-Plagwitz (2008) explores the pros and cons of Second Life as a tool for language education. Akayoğlu (2011), uses Second Life to analyse social language of EFL learners in 3D platforms. As it can be seen, CALL is a field that is improved by the tools that was developed using computer technologies and adapting those tools to language learning and teaching.

Thus, there is a brief history of how CALL became CALL as used and known nowadays. But just as was mentioned, new tools, new opportunities further enhance CALL. For instance, Web 3.0 also called “semantic web” has emerged by the end of the first decade of 2000s (Naik & Shivalingaiah, 2008). As Fırat and Fırat (2020) described “Web 3.0 is called semantic web because it creates meaningful information”. The prominent features of these technologies are; smart web with intelligence analysis, personalization, interoperability web, virtualization and multimedia (Wadhwa, 2015). Rajiv and Lal (2011) defined Web 3.0 E-learning tools as; semantic web, augmented reality, intelligent tutoring system, 3D visualisation, online 3D games, 3D-Wikis, virtual worlds, and virtual reality that provide personalized settings for individuals.

Nowadays, perhaps one of the biggest breakthroughs in the computer technologies has changed the world. An AI model was introduced by OpenAI that is known as ChatGPT. “ChatGPT is a state-of-the-art language model that has revolutionized natural language processing by generating human-like text with context and coherence, enabling new possibilities for human-AI interaction.” (Roumeliotis & Tselikas, 2023) Artificial intelligence has been slowly overtaking the internet and AI tools have been dominating the related fields. After ChatGPT, many chatbots were introduced, BardAI, NotionAI, and BingAI, are just a few of them. AI enhanced tools started to reshape the field of language learning and teaching and CALL studies nowadays are mostly centred around the biggest axiom of the 21st century so far, that is “artificial intelligence.”

1.5.5 Artificial Intelligence and Language Learning

Artificial intelligence (AI) has been one of the main research topics of computational sciences for over half a century. Long before its recent uprising AI had already been utilised in many if not the most technological tools that we use on a daily basis. Everything from phones, computers to personal assistants, electric cars or even apps we use like Instagram, Spotify, X (formerly Twitter), YouTube have been using AI to make our experiences of using them easier. Ad or product recommendations, spam messages, GPS and many more are results of AI mechanisms such as deep learning or machine learning that will be explained in detail. Whilst that is the case it is natural to ask the most basic question, what is Artificial Intelligence? “made by people, often as a copy of something natural” (Cambridge University Press, n.d) is the definition of artificial and as for intelligence we see the definition as “the ability to learn, understand, and make judgments or have opinions that are based on reason” (Cambridge University Press, n.d.). According to the definitions of “artificial” and “intelligence” it can be deduced that artificial intelligence is the copy that made by humans to mimic or replicate the ability to learn, understand and make judgements like humans do. This is the most basic explanation of AI but not the most complete. Russell and Norvig (2010) divide the explanation of AI into four distinct parts as follows, **Acting Humanly, Thinking Humanly, Thinking Rationally, and Acting Rationally.**

In 1950 Alan Turing, the Father of Modern Computers, proposed a test known as **The Turing Test**. The aim of this test was to propose an effective description of intelligence. Test is conducted by human interrogators who try to

identify if the interrogatee is human or computer via written responses to their written questions. A computer passes the test when the interrogator thinks the answers are from a human interrogatee. To that a computer needs to have four fundamental abilities; **natural language processing** to communicate effectively in English language, **knowledge representation** to collect knowledge and information, **automated reasoning** to effectively use the information to answer questions and deduce new assumptions, **machine learning** to adjust itself to new environments. Moreover, since the test included videos and objects, it also needed **computer vision** to observe objects and sometimes **robotics** to handle them. Thus, when inspected the issue at hand for acting humanly AI can be explained as “The art of creating machines that perform functions that require intelligence when performed by people.” (Kurzweil, 1990).

One other explanation of AI as mentioned before is its’ ability to think like a human being in other means **The Cognitive Modeling Approach**. In their book Russell and Norvig (2010) dives into the issue of “think like a human” as follows, “If we are going to say that a given program thinks like a human, we must have some way of determining how humans think. We need to get inside the actual workings of human minds.” They suggested three ways to get inside human minds. The first way is possible through introspection, meaning “...trying to catch our own thoughts as they go by...” (Russell and Norvig, 2010, p3). The second way is monitoring people in action in other words psychological experiments. Finally examining the brain in action via brain imaging. It was claimed that after we understand how human brain thinks, we can apply the theory behind thinking to computers. **Cognitive Science** is an interdisciplinary field that tries to create complete theories on how human mind works by AI based computer models and experimental psychological techniques. From this perspective Bellman (1978) tries to explain Artificial Intelligence as follows; “[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning ...”

The third modal for explaining AI is **the Laws of Thought Approach** or **Thinking rationally**. This modal is not only a question for Artificial Intelligence but also it is a quest for human minds as well. Acknowledgement of this issue dates back to ancient times. Aristotle tried to systemise correct way of thinking in his book “Prior Analytics” and defined **syllogism**. Syllogism is a logical argument that

uses deductive reasoning. Deductive syllogism occurs when two true premises indicate a conclusion. A famous syllogism that was associated with Aristotle was “All men are mortal, Socrates is a man, therefore Socrates is mortal”. This quest for searching the laws of thought throughout the time starting from ancient times with Aristotle and The Stoics like Chrysippus in the West, Master Mo, Hui Shi and other figures in the “name school” in the East, later followed by medieval philosophers such as Ibn Sina, Thomas Aquinas, Jean Buridan and many more led to the field of **logic** in the mid-nineteenth century. With the development of this field, and computer sciences, thus AI, Artificial Intelligence has been heavily influenced by logic and logical ideas. Even though the value and relative importance is questioned by some leading specialists, many agree that logic plays a critical role in at least specific central areas of AI studies. Looking at AI from this perspective Charniak and McDermott (1985) comes up with this explanation for it; “The study of mental faculties through the use of computational models.”

As for one of the last fundamental explanations, **The Rational Agent** approach meaning **Acting Rationally** welcomes us. First of all, the meaning of an **agent** must be understood. An agent is something or someone that does the action. An acting participant of an equation. Unlike people who are supposedly free to act or not, a computer program must do something, but computer agents must do more such as autonomous operation, high perception skills, persistence, adaptation, and evolution and so on. Since AI is a computer agent, all AI tools must do a task. However, what separates AI from other computer programs is that it must be a **rational agent**. The difference is that a rational agent is expected to act in order to reach the best possible outcome or the best expected outcome (Russell and Norvig, 2010). The rational agent approach is the most comprehensive approach to Artificial Intelligence. In this approach, AI programs are expected to use the “laws of thought” every now and then because reasoning and using logic is one important way to achieve the best possible or expected outcomes. What is different this time is that sometimes there is no “right” way to do something, but you still must act. This approach tries to cover that as well with the agent hypothetically being able to find optimal opportunities. Also, an agent needs all the skills from the Turing Test to act rationally. Moreover, this approach has two advantages over the other approaches (Russell and Norvig, 2010). First of all, it is a lot more wide-ranging and secondly it is more open to scientific development to models solely based on

human mind. Regarding this approach an explanation as such can be made “Computational Intelligence is the study of the design of intelligent agents.” (Poole et al., 1998, p.1).

1.5.6 Interdisciplinary Nature of Artificial Intelligence

Now that the general understanding of what an AI is given, this section will be devoted to exploring some of the fields that have contributed to the development of AI. Since the idea of AI was first developed, many questions arose in many fields to further investigate the issue and produce solutions to the problems around AI. One reason for this quest is the goal to create thinking, acting agent as aforementioned. In order to achieve that goal fields like **economics, mathematics, philosophy, linguistics, neuroscience, psychology, computer engineering**, and many more has contributed to AI by asking questions and seeking answers to foster improvement. Nilsson (2009) explains this process of asking questions as such, “Clues about what might be needed to make machines intelligent are scattered abundantly throughout philosophy, logic, biology, psychology, statistics, and engineering.” Though most of the fields seems irrelevant to each other, they all have one similar trait; that is to try to understand humans and living creatures in some aspect. How does this benefit the study of AI? It is a necessity to understand humans if we want to create better, faster, stronger, and more intelligent versions of us.

Philosophy is one of the fundamental steps of understanding humans. It is one of the oldest disciplines in history. Teichmann and Evans (1999) describes philosophy as a “study of problems which are ultimate, abstract, and very general. These problems are concerned with the nature of existence, knowledge, morality, reason, and human purpose.” Initially, philosophy or Greek version philosophia (φιλοσοφία) literally translates into “love of wisdom” (Webster’s New World Dictionary, n.d). It is a multifaceted subject and has evolved into a broad concept throughout history with multiple sub-disciplines such as rationalism, dualism, empiricism, and many more. With its interdisciplinary nature, philosophy deals with the fundamental understanding of subjects. The importance of philosophy in terms of artificial intelligence is the quest of trying to comprehend the human mind and the idea of knowledge. Questions such as “What are the differences between the human mind and brain?”, “What is the origin of knowledge?”, “How does the knowledge leads into actions?” are crucial for science to understand the human

mind, and apply that information to create better, more successful, and smarter AI tools.

One other step in identifying humans is **psychology**. Recognition of psychology as a science is rather recent and dates to mid-19th century. Although it is not an ancient science, human psyche was studied long before it became a science. According to John Dewey (1892) psychology is the science of the facts or phenomena of **self**. From the earliest times the definition and applications of psychology have changed slightly, however even now psychologists ask similar questions to the questions that were asked thousands of years ago. “How do humans act?” “Are humans born with knowledge?” “How do humans think?” These are some of the questions which are asked to understand human behaviour. Once it is understood, human behaviour could affect how the AI tools would act, think, and behave. This goes without saying that it would improve Artificial Intelligence in terms being more human-like.

Artificial Intelligence has taken many appearances and many forms. In daily life, it is impossible not to encounter with tools such as image generators, chatbots, advertisements, and so many other examples of AI tools. Some AI tools such as chatbots, writing tools, and translation tools are required to act human-like to give better results. Therefore, **linguistics** is a critical field to consider in the development and improvement of certain AI tools. Lyons (1968) explains linguistics as the scientific study of **language**. Language is one of the most complex tools humans have used. It is a long process to explain language in detail, but Kracht (2007) briefly describes language as a means of communication. Language recognition, language production, language processing, and language comprehension is vital for specific AI tools. Exploring human languages and trying to enlighten the process or the reasoning behind languages could help us to create more effective and reliable AI instruments.

Even though the number of fields that are related to AI could be explored more with at least one field for a specific purpose, in terms of language education and other educational means, understanding human behaviour and mind are one of the main purposes. Other studies on AI in a different field in education or one that is different altogether can shed light on diverse areas of science and explore the relationship between AI and them. Consequently with the aspects that were defined

above, lately AI has been covering an important number of research in language teaching, learning, applied linguistics, and language studies.

1.5.7 Artificial Intelligence and Language Learning

Therefore, Artificial Intelligence, as was explained, slowly but surely taking over every field. One main reason for that is AI enhanced tools completes tasks a lot easier, and more efficient. Even this is a legitimate reason to turn everyone's attention towards AI. Teachers are able to create modern solutions to old problems. ChatBOTS or AI enhanced tools can prepare detailed lesson plans, long-term goal-oriented and personalised study programs for learners, and certainly not least grade students based on their performance. ChatGPT, an AI chatbot published by OpenAI by the end of 2022, has changed the way AI is included in language learning and teaching. However, long before that there were already AI enhanced tools, application, programs, and assistants used by learners and teachers to learn or teach a language. For instance, as Ali (2020) mentioned, AI assistants such as Siri and Cortana were effective tools for ESL or EFL learners to practise their English, "... it helps the group of learners to improve their speaking; in particular pronunciation and listening skills." There were, and still are, applications such as Duolingo (AI enhanced language learning program) that created personalized lessons and study plans for learners, which effectively used authentic materials and aimed to increase learners' exposure to target languages. In conclusion, digital prints of AI are not new to the field of language learning and teaching, but every other day its impact is exponentially growing.

Pokrivcakova (2019, p. 137) explains AI's status in linguistics and language teaching as follows: "Using AI and NLP (natural language processing) help create more detailed descriptions of natural languages, leads to better-processed corpora, as well as to a better understanding of mental processes occurring in human brains while verbally communicating."

It has been evident that employing AI as a supporter in language teaching and learning proved to be useful to develop language specific skills, such as reading comprehension, practicing repetitive language skills (N.Y. Kim, 2019), correcting English pronunciation, also as Pokrivcakova (2019) showed, it can help with broader activities, comprising automatic marking, providing feedback, adaptive educational experiences, intelligent tutoring and prediction systems (Crompton et al., 2024). Pokrivcakova (2019) divides the uses of AI into three categories in

education, first one is “learner facing” that are used by students to learn by themselves, second one is “teacher facing” used by teachers to aid in teaching activities, including grading or planning, and the last one is “system facing” used by administrators to control student data. To answer the question of “Why use AI?”, AI technologies are efficient whilst creating user specific programs, they can provide plans that match individualistic goals. “AI can mine large amounts of data, operate using natural language (speech, listening and writing) and follow rules and patterns of language.” (Crompton et al., 2024)

In the past, CALL technologies were human-dependent, meaning that in order to create programs, applications, and practices, teachers needed to provide all the “footwork”. Even though it required a bit of work, it was a lot simpler and easier than creating handmade materials. As it was discussed, CALL studies made teachers’ lives considerably more comfortable. However, there is something newer that will further improve the quality of the work for all stakeholders in learning a language. “The shift from CALL to ICALL (Intelligent CALL) has been inevitable and brought substantial change in the quality of student-computer interaction.” (Kannan & Munday, 2018) The meaning of Intelligent CALL is basically a system that can provide valuable advantages to the field of learning and teaching a language by being able to systematically act like a human and create better and improved versions of learning environments with relatively less effort. Pokrivcakova (2019) explains one of the expected benefits of ICALL stem as a huge potential which AI can provide to make digital language learning truly personalised to each learner. As it was explained in the Artificial Intelligence section of Literature Review part, most of the current AI tools that are commonly used have profound machine learning algorithms, and they are able to process nearly endless amount of data which adapt in real-time according to the user behaviour. They can analyse each user’s strengths and weaknesses and produce personalised set of materials in every session. As for teachers, they are able to effectively respond to the teachers’ needs whether they ask to evaluate an essay or help creating an entire lesson plan for a group of learners. Furthermore, Campbell-Howes (2019) indicates something at least as important, which is the algorithm is able to learn from individual and collective learner, teacher behaviour that eventually makes its predictive power even stronger. Moreover, the more users use, benefit, and provide data to AI tools, the more the tools learn and process entirely new datasets each time. One other benefit of ICALL as

Pokrivcakova (2019) included; learners are able to have their own pace of progress, there is instant feedback as a strong motivational factor, individualized repetition of topics and emphasizing activities where a learner has had weaker output, quick and objective assessment of progress, better comprehension of learner needs, preferences and learning strategies, being able to successfully predict learners' future performance. Teachers can use AI as a consultant and helper whilst learners go all the way benefiting from AI as mentioned. Of course, needless to say, this is possible by shifting CALL studies into ICALL studies, including AI research to further improve possibilities and making language learning and teaching easier for both teachers and learners.

To effectively do that, relevant areas must be understood. Although all language studies and all AI studies are relevant to each other to a degree as Schulze (2008) pointed out highly relevant areas of AI research for CALL studies include natural language processing (NLP), user modelling, expert systems, and intelligent tutoring system (ITS). One important tool that provides us with the necessary instruments are certainly ChatBOTS. Intelligent chatbots have caught the attention of language educators because they can interact with learners in their target language instantly and realistically (J.H. Lee et al., 2020; Kohnke et al., 2023; L. Fryer et al., 2020;) For the reason that chatbots are computer applications that simulates human-like natural conversations with users through text and/or voice (Ashfaque et al., 2020) also they can offer synchronous support and tutoring (Kerly et al., 2007) State-of-the-art chatbots are leveraged by NLP, deep learning, and machine learning technologies and the reason they are effective, they learn how to respond to user inquiries based on a vast human language dataset (Jiang et al., 2022) Therefore, they can communicate with users intelligently, continuously learn from prior interactions, improve over time, and act as a tireless language-learning assistants (L.K. Fryer et al., 2019; N.Y. Kim, 2018) Rich input is essential for successful second language acquisition (Kohnke et al., 2023). W. Huang et al. (2022) indicates that AI-powered chatbots are able to provide linguistic input and daily conversation practice. Consequently, practicing with improved chatbots might help learners to get a grasp of daily use of the targeted language, and they can practice their language skills without anxiety, pressure and with immediate feedback if needed. Furthermore, chatbots are likely to stimulate language learners' interest (Gallacher et al., 2018; Kohnke, 2022b; L.K Fryer et al., 2017), contribute

to their general growth in various language skills (H. Kim et al., 2019; Kohnke, 2022b; N.Y Kim, 2019), additionally since chatbots are not real humans and available at anytime and anywhere, learners can practice at any given time and any place they are (Haristiani, 2019; Winkler & Soellner, 2018). When learners find the input challenging, chatbots can easily adjust the level they are speaking matching the proficiency level of the learners (Kuhail et al., 2023). Chatbots can offer a range of the use of language with a variety of expressions, questions, vocabulary items, that human partners may not (W. Huang et al., 2022), and assist in providing an authentic, interactive language-learning environment.

Teachers can also get at least as much help as learners from chatbots. AI-enhanced chatbots, with enough dataset and accurate prompts, can effectively conduct formative assessments and provide immediate feedback (W. Huang et al., 2022; Kuhail et al., 2023). Furthermore, teachers can differentiate materials for their students and regarding their needs using chatbots (Kohnke et al., 2023). With a reading passage, which can also be generated by chatbots, teachers can ask chatbots to prepare reading comprehension and expansion questions, after providing enough data about a group of learners chatbots can generate exams to assess specific language skills and evaluate those exams providing useful feedback.

Crompton et al. (2024) showed AI-enhanced tools, whether it being chatbots or other AI programs, are used in six groups in English Language Teaching (ELT). Although there are overlaps in using AI, their research suggested that the main codes are *speaking, writing, reading, pedagogy* and *self-regulation*. This grouping is important to understand the main reasons that AI programs are used in CALL fields and what benefits are gathered from using AI tools as assistants. It is worth noting that those codes will help to determine the type or types of AI tools to use whilst addressing specific needs and/or language skills of learners, and the purpose, aim, and needs of teachers in order to help them ease their lives by making repetitive, administrative task easier and to create efficient and effective learning environment for both learners and themselves as well.

To begin with, studies revolved around speaking mainly focused on pronunciation skills (Crompton et al., 2024). EFL learners generally face with challenges in speaking due to the lack of interaction with English outside of the language courses (Tokoz & Goktepe 2014). AI can be a helpful assistant to deliver that interaction between learners and the English language. In a study conducted by

S. -C. Liu and Hung (2016) the AI drastically improved students' pronunciation skills by focusing on pitch and intonation patterns. They found that the visual illustration by AI aided students to notice patterns and support their pronunciation skills. Dizon and Tang (2020) used Alexa, a personal AI assistant by Amazon, and asked students to have conversations with Alexa. According to the students, Alexa was easy to use, and it stimulated meaningful conversations, supported vocabulary acquisition, improved language skills, and provided interesting, enjoyable learning experiences. "AI has a direct connection between speaking a language and natural language processing in that AI can recognise and produce speech." (Crompton et al., 2024). Kazu and Kuvvetli (2023) developed an AI-enhanced pronunciation program. The system was developed to aid students whilst practicing pronunciation, by recording and reacting to students. The results indicate that the vocabulary items remained in memory longer and learners mostly benefited from learning consonant and vowel sounds.

Additionally, the newest model of ChatGPT, that is ChatGPT 4-o offers a strong version that is able to have human-like spoken conversations with users. ChatGPT 4-o carries the text-based written interactive form of ChatGPT to another level where users can actually speak with the enhanced chatbot. Although there is not any research in literature since it is just released at the time of this thesis, it is a technology that is certainly worth looking for. It is natural to suggest that the introduction of ChatGPT 4-o will affect this area of speaking opening wider areas changing the way speaking AI assistants are perceived around the globe, and it is a huge opportunity for researchers to investigate it.

As for the writing skills, the focal subskills were vocabulary learning and grammar (Crompton et al., 2024). Lo (2023) investigated vocabulary improvement and retention in writing of students when they had access to neural machine translation programs. It is found that AI programs helped students to expand their vocabulary knowledge particularly when specialised or unambiguous expressions were introduced.

Another common AI tools used in writing are grammar checkers. Various disciplines use grammar feedback tools such as Grammarly (Koltovskaia, 2020). Dizon and Gayed (2021) explored the effects of Grammarly in higher education in ELT. They discovered that students who used Grammarly had less grammatical errors and had a variety of lexical items than students who did not use any AI

enhanced grammar checkers. Recommendations by grammar checkers such as real-time corrective feedback supported students studying English. Nazari et al. (2021) studied Grammarly as well and the results were again positive. It is shown that there was a considerable improvement in behavioural engagement, emotional engagement, cognitive engagement, and self-efficacy in writing.

Generally, translation tools are seen as a negative aspect of technology and/or AI. However, Chon et al. (2021) showed that using Google Translate helped students to generate the largest number of complex words. It also allowed learners to procure a greater range of lexical items as the translation tools led students to get exposed to the different lexical choices available to them to effectively express their intended message.

The next code is reading and although it is not as common as speaking and writing, reading was one of the codes of AI studies as well. The reason might be that AI algorithms such as NLP, or machine learning are aligned to speaking and writing that researchers want to use (Crompton et al., 2024). It is found that the only focus was on vocabulary skills in reading studies. Zheng et al. (2015) examined how vocabulary learning in reading happens through gaming quest-play mediated in English with Japanese player playing with an English player. Using avatars, the learners used semiotic resources infused in World of Warcraft (an MMORPG game). For instance, in certain occasions players see the word 'looting' appearing. The students learn that word via interaction with other players, then they are able to use that word in other contexts as well. Zheng and colleagues suggest that students have ample opportunities in gaming beyond what textbooks or classrooms can provide. "For educators who have pupils who appear unengaged or unmotivated to learn, gaming could be an approach that supports those pupils in reaching their learning goals." (Crompton et al., 2024, p. 2518).

The code pedagogy is used to identify the methods, strategies, and techniques used to facilitate ELT and they were often connected to language skills (Crompton et al., 2024). Kim (2022) inspected the impact of the pedagogical approaches of score predictions, lectures, explanations, and practice tests on Korean students. *Soljam* which is an AI system was used to aid students. Students underwent a diagnostic assessment then AI used the students' data to prepare lectures, explanations, and tests specific to each student. Lee et al. (2023) explored a context-based approach; learner-generated context based in particular. The

context is generated from students' data as they performed activities and made choices. Then the system learns about each student to show content that aligns with their preferences. It is shown that the LGC AI approach facilitated autonomous learning experiences. With the help of ever-growing field of AI, researchers and teachers can benefit from the advantages that AI has to offer to create different learning experiences. The last code that was found in AI studies is self-regulation that "...refers to the ability to manage and control one's thoughts, emotions, behaviours, and physiological responses to achieve personal goals and maintain well-being." (Crompton et al., 2024, p. 2520) Although there are some concerns about students being over-reliant to the AI technologies and avoid mastering contents themselves, some studies indicate that AI enhanced tools that are used to foster self-regulation might help students to become more autonomous learners, set more effective goals and mitigate the effects of anxiety of using English. (Crompton et al., 2024; He, 2021) A study conducted by Hew et al. (2023) showed that students that used chatbots who lacked self-regulation in online language courses feeling disengaged and isolated, got significant help from chatbots whilst clarifying their learning goals, developing techniques for setting goals and raising awareness of learning strategies in goal setting. It is also reported that students enjoyed using chatbots and had a positive learning experience. Chen et al. (2022) explored robot-assisted language learning. Chen brought AI and VR together to create a system to use robots for training English language tour guides.

"Pupils played the role of tour guide or traveller and interacted with various exercises including proper working, matching words, collocation and sentence structure. As the pupils worked with the AI system, the findings of the study identified affordances including developing autonomy, while also increasing motivation and engagement." (Crompton et al., 2024, p. 2521)

Anxiety was another focus of the code. In 43 different study, researchers used AI to lessen the effects of anxiety as students attempted to grasp another language and it is often referred in the findings that AI helped reducing the anxiety. In their studies Chen et al. (2022) and Çakmak (2022) discovered that AI helped to improve students' skills and played a role in reducing their anxiety. With all the data gathered, it is supported by evidence that AI technologies help students and teachers in more than one way. AI is redefining education; language education is specific. The roles of teachers are accordingly changing to adapt this evolved era of

technology and artificial intelligence. “Even though human teachers and social interactions beyond the digital environment are still essential for mastering a second language, the application of CALL into foreign language learning, including the elements of AI, leads to redefined roles for teachers and learner” (Lam & Lawrence, 2000; Pokrivcakova, 2019). First of all, it is worth mentioning that preparing future EFL teachers for ICALL is vital. It is argued that teacher attitudes, beliefs, and preferences play a critical role in adapting CALL methodology (e.g., Albirini, 2006; Atkins & Vasu, 2000; Beatty, 2003; Cummings Hlas et al., 2017; Egbert, Paulus, & Nakamichi, 2002; Goertler & Winke, 2008; Hong, 2010; Hoven, 2007; Hubbard, 2004, 2008; Hubbard & Levy, 2006b; Kessler, 2006, 2007, 2010; Lam, 2000; Levy, 1997; Liu et al., 2004; Lord & Lomicka, 2011; Luke & Britten, 2013; Peters, 2006; Pokrivcakova, 2019; Pokrivcakova, 2015; Straková & Cimermanová, 2018; Williams et al., 2014) That is why it is important for EFL teachers to have an appropriate training to include AI technologies in classrooms. Main focus should be to help them feel well-prepared and confident to use AI technologies in the new era of teaching (Pokrivcakova, 2019). It is only then teachers can adapt to the changing role of teaching a language. Contemporary technologies allow students to be self-autonomous now more than ever. Students are facilitated to be in control of their learning process, choosing the pace themselves. Accordingly, teachers should abandon their previous roles of being the only authority and decision-maker to be a facilitator and supporter of learners (Banerji, 2006; Rilling et al., 2005).

Regardless to say, although AI technologies are in a position of power, it is crucial not to be over-reliant on AI tools. Ubiquitous presence of AI brings some considerations with it such as ethics, data privacy, data protection, quality of language learning and teaching, resistance to change especially by teachers, overreliance on technology, cost, accessibility and more. However, these challenges and considerations are surely will be overcome by rapid evolution of AI and technology, also the frequency of AI technologies used by teachers will help in this process. So, instead of alienating the new, it is beneficial to adapt to the new. Because technology, AI, and even students are evolving every other day, so must teachers. Even though the AI technologies are rather new on this scale, it is needless to say that it is a matter of time before they become the mainstream in language learning and teaching.

1.5.8 CALL and Writing

Besides AI studies in the field of English language learning and teaching, writing has been quite a popular area of study for CALL researchers. It is widely accepted that writing is an important skill and arguably it is difficult to learn. That's why Turmudi (2017) suggests that teachers should have proper tools to facilitate students' enthusiasm and keep them motivated in terms of writing education or writing related tasks. Computers are one of the many tools that can help us to achieve that goal. As Jafarian et al. (2012, p. 139) put it:

Computer-assisted writing instruction proposes an alternative method to the traditional ones. This approach may overcome some shortcomings of traditional methods of writing to some extent by providing feedback about students' mistake /errors, along with some explanations and suggestions. The feedback is a mechanical, grammatical and other formal language elements inaccuracies, which are assumed to affect one's writing.

Among many reasons, this is only one important reason to use the effectiveness of computer technologies in language learning and teaching. Grabe and Kaplan (2014) indicated that teaching and learning practices that include self-discovery, multiple drafts with feedback between drafts, and feedback from peers, teachers, or other formative assessment can bring a more process-oriented approach in writing. (Jafarian et al., 2012) Computers give students, learners or even teachers the chance to receive feedback from multiple sources. Even more, computers are able to give immediate feedback even when in the process of writing a text, an essay or a document via correcting the wrong structures, suggesting different lexical items, adjusting the wrong pronunciation of words, and overall making improvements on the texts.

CALL studies around writing initially had focused on two major areas which are developing word processing skills and the use of text-based or graphic organizers to support the writing process (Zaini & Mazdayasna, 2014). According to Hirvela (2005) it is necessary for both readers and writers to learn about the world of technology and at utmost urgency to introduce them to the screen culture since college writing is becoming computer-based and computer is used as an effective, popular writing tool. According to the study of Slattery and Kowalski (1998) there were two trends in CALL, the first one started in 80s until the early 90s which was using word processors to improve writing quality and raising motivation of the learners because they were presented a new way of writing and typing, and the second one was computer-mediated communication (CMC) surfaced with the

introduction of the internet and the role of hypertext ...*which link a related text to numerous texts* (Zaini & Mazdayasna, 2014). In their study, J. Li and Cumming (2001) aspired to investigate if word processors could affect the writing process and improve the quality of writing of second language learners. Learners who used computers got higher scores of their tasks. “Word processors help reduce the mechanical difficulty involved in changing texts and offer a fluid and easily transformed communication, users might create longer compositions and do more revisions of their writing than they would do with pen and paper.” (J. Li & Cumming, 2001, p.128) A study by J. Li (2006) examined the effects of word processors on ESL students. It is found that participants were more focused with their higher order thinking activities on a computer, and they made revisions significantly more on a computer than on a pen and paper. He suggested that educators should consider the obvious influence of computers on writing assessments (Zaini & Mazdayasna, 2014). Also, another study by Zaini and Mazdayasna (2014) inspected the impact of CALL on the improvement of EFL learners’ writing abilities. Study conducted with two student groups, one group was asked to complete writing tasks on a computer and given computer-based instruction, whereas the other group took a traditional approach to writing. Results bared that computer-based instruction and the students that performed the tasks on computer “... outperformed their counterparts in terms of using appropriate articles and tense.” Researchers claimed that the automatic feedback by Microsoft Word Office besides the recommendations of instructors, helped them to improve their common sense, linguistic competence, and logic to correct themselves. There have been numerous studies that proved CALL to be quite useful in writing. Overall advantages of CALL in EFL teaching were found out to be that CALL can improve academic performance, enhance motivation, and promote learning. Moreover, aligning with previous studies, it is shown that CALL technologies allowed students to get higher grades in writing tasks. In addition, considerable amounts of studies revealed that EFL students had positive attitudes to CALL technologies in teaching and learning writing. To conclude, literature around CALL and writing highly emphasizes that even when there are no significant differences in terms of achievement between students that use computer than those that do not use computer in writing, using CALL would still help students to increase their motivation and self-efficacy. Furthermore, researchers suggest that CALL studies

in writing helps students to self-evaluate, resulting in better understanding of the English language in the long-term. With the evolution of CALL technologies to ICALL, there are a lot better, faster, and more improved tools that can be used in writing nowadays. Consequently, the relationship between computer technologies and writing in English will be more intertwined than ever especially after the recent AI technologies.

1.5.9 Artificial Intelligence and Writing

Writing has always fascinated the humankind and was considered as one of the most unique skills and one of the many ways of displaying human intelligence. Invention of writing has changed the world dramatically. It allowed people to express themselves and communicate in another way than speaking, also it further improved human brain resulting in abstract thinking (Olson, 2009). Moreover, writing is a form of time-travelling, breaking the barriers between the past, present and future. Science is what it is now thanks to writing. All in all, writing is one of the strongest tools that people have since the Sumerians invented the first known form of writing in about 5-6 millennia ago. (Goody, 1986) Although, if the writing is described as a “visual communication” tool, the date could be extended to somewhat 40.000 years (Olson, 2009). And for a few decades, researchers in AI and NLG (natural language generation) have been trying to reinvent the wheel, or writing in this case, to achieve that evolution of human intelligence for AI technologies (Dehn, 1981; Klein et al., 1979; Liu & Hung, 2016; Meehan, 1977; Ippolito et al., 2022).

Artificial Intelligence is changing everything we do including any form of communication. (McKee & Porter, 2020) Additionally, AI has become a popular area of interest for researchers in Education and studies intertwining both have been increasing exponentially. Specifically, NLP is a sub-field of AI that was used by educators in particular by language educators. (Chen et al., 2022) A study by Alharbi (2023) examines the types and number of AI tools that can be used and are used which focuses on writing by learners and teachers in EFL settings. The study divides AI-powered writing assistance technologies into four; (1) Automated writing evaluation (AWE) tools such as Criterion, and WriteToLearn, (2) Automated writing corrective feedback (AWCF) tools such as Grammarly, and ProWritingAid, (3) AI-enabled machine translation tools including Google Translate or other online dictionaries, lastly (4) Automatic text generation and Deep

Learning Technology. These tools are some of the most common AI-enhanced writing tools, and assistants for now. There is no doubt that both the tools and the fields will expand in the near future. The reach of AI is growing every other day and educators, teachers should help students to guide which of these tools to use and how to use them in different scenarios. For instance, Kirschenbaum (2017) explains that digital writing has been an area of interest to applied linguistics and second language acquisition studies since 1980s. Digital writing is an umbrella term that covers a variety of skills which are successful online interactions, navigating social networking sites, blogging, online communication, and writing using word processors. A study by Moore et al. (2016) showed that digital writing tools can respectively improve the writing quality of students, and it is crucial to ensure a complementary face-to-face guidance to digital tools by educators. Another study by Homouma and Menezla (2019) pointed out the strong positive correlation between academic writing performance and digital literacy on EFL students. Furthermore, Purcell et al. (2013) explains the positive influence of digital technologies on native-speaker students' writings.

One important function of AI-based technologies are text predictions. Algorithms that AI-tools use such as deep learning, machine learning and NLP allow text prediction, which helps users by completing their sentences according to the words they used, and the frequency of words that were previously used in conversations. Especially in recent years, text predictive assistants have gained popularity, and companies try to integrate text prediction to make users' lives easier. Gayed et al. (2022) conducted a study using an AI-based tool, "AI KAKU" to observe the impact of AI KAKU on writing. Although the results did not show significant effect, the researcher suggested that AI KAKU is potentially a lot more useful once more assistance is provided to English language learners. "Predictive text and AI-based learning agents as a whole are a growing trend in education." (Fung, 2010; Waldron et al., 2017)

Studies around AI and writing skills further expand to even academic settings. Khalifa and Albadawy (2024) identified six domains that AI is used in academic writing and these are; (1) Idea development and research design, (2) content development and structuring, (3) literature review and synthesis, (4) data management and analysis, (5) editing, review and publishing support, (6) communication, outreach and ethical compliance. They suggested that academic

institutions and researchers should include AI tools in their research to benefit them, but whilst doing so they must consider ethical integrity and their research should be original. Their research indicates that AI is an important tool in academic writing settings.

AI transforms idea development and research design by providing valuable insights and optimising methodologies. It enhances content quality through writing assistance and emotional tone analysis. In literature review and data management, AI's ability to process large data sets ensures comprehensive analysis and integrity. AI also streamlines the publishing process and supports ethical compliance in research dissemination." (Khalifa & Albadawy, 2024, p. 10)

It is no doubt that digital writing tools, regardless of the setting, support the quality of English writing as evidence suggests (Nobles & Paganucci, 2015). AI-enhanced writing tools seems like will support writing journeys and probably cover the most of it in the near future. This much reliance, use and perhaps overuse of AI raising some ethical concerns as well especially in terms of communication ethics (McKee & Porter, 2020). Although the ethics from academia to learning and teaching the English language is a question, it is needless to say that AI-enhanced writing tools and AI technologies will surely become the biggest helpers of researchers, teachers, and learners. Whether it is a desired outcome or not students are most likely to use text generators, writing assistants, chatbots and more whenever they find the chance (Otsuki, 2020). It is in everybody's best interest that in every step of the learning and teaching process teachers too must be familiar with AI technologies from K-12 education to university education. Especially in EFL settings English language teachers, and English language teacher educators must be aware of the fact that AI is improving every other day. Embracing the facts that cannot be changed will surely improve teaching experience as well as learning experience for students.

1.5.10 CALL and Feedback

"Issues regarding the role and contribution of corrective feedback for language learning have been central to second language acquisition (SLA) theory and pedagogy." (Heift, 2004) Feedback is explained in SLA as a post-task information given to learners regarding their language performance on an assigned task by generally teachers, but recently programs, or applications (Spada & Lightbrown, 1993; Ellis, 1997; Doughty, 2001; Liutkus, 2012; Duhon et al., 2015). Feedback is considered to be a crucial part of language learning for learners from

K-12 education to college, university level education (Fajfar et al., 2012). The most important aspect of feedback is to help students realise their performance and the performance that is optimal (Kang & Han, 2015). Another purpose is to scaffold learners on their journeys to the desired level of performed language (Ilies et al., 2010; Meo, 2013, Mohamed, 2021). Scaffolding, which is the support of a knowledgeable person to the less knowledgeable person to complete an objective that they are unable to do themselves, is critical to strengthen learning outcomes (Kargar & Tayebipour, 2015; Masuda & Angela, 2015). According to Van Lier (1996) feedback is a classroom discourse. It is the evaluation of the Initiation-Response-Feedback (IRF) model. SLA research show that teachers have several ways and strategies to initiate feedback process (Allwright, 1975; Chaudron, 1977; Seedhouse, 1997; Lyster & Ranta, 1997). Sociocultural theorists suggest that students should not be left alone and have to work without any assist and guide for language learning, additionally they point out the importance of co-participation in learning for students to produce a “common product or artifact” (Ai, 2017; Anh & Marginson, 2013; Dongyu et al., 2013; Elola & Oskoz, 2016; Englert et al., 2006, p.209; Hung, 2016). Given feedback can vary in a number of ways changing from context to context and student to student (Ferreira et al., 2007). Furthermore, it is found that the efficacy of the feedback is affected by several factors (Heift, 2004). In their study, Havranek and Cesnik (2001) indicated that successful corrective feedback is influenced by its format, the type of error, learners’ attitude towards feedback, proficiency level and the verbal intelligence of the learner are all considered to be an important aspect that determine the effectiveness of feedback. In addition, Ferreira et al. (2007) suggested that corrective feedback is “...most beneficial when it occurs in response to naturally-occurring errors or in the context of ongoing efforts to communicate.” Overall, feedback are important tools that teachers hold that can enhance learning and facilitate student motivation, interaction and engagement. Feedback, whilst it is important, is extra work for teachers. Furthermore, especially in open-ended tasks, teachers may struggle to effectively evaluate and give feedback to students. Noticing the extra labour, researchers have been incorporating feedback into CALL studies. With the informative and helpful tools that computers possess which can successfully analyse language, many programs have been used to provide feedback.

ITS, for instance, have benefited from the NLP technologies to analyse natural input from students (Livingstone, 2012). Kreindler (1998) claims that there are a few criteria for CALL tools to give good feedback: “(1) focus on the content and meaning, (2) support learning instead of testing, (3) be communicative and moderate, (4) get personal involvement, (5) promote cognitive skills, (6) provide cultural enrichment, (7) differentiate among students, (8) be simple clear and economical” (Livingstone, 2012). It is safe to assume that the criteria and the ways of giving and receiving feedback have changed in the last quarter but the general idea of Kreindler (1998) that feedback is continuous and important part of the learning in a mostly cognitive way remains similar. Effectiveness of CALL assisted or CALL dominant feedback have been researched since 1980s. Robinson (1985) was one of the first studies that showed to usefulness of CALL feedback. Nagata (1993) showed that students who received “intelligent” feedback had a significantly higher rate of performance than students who did not have that feedback (Nagata, 1993; Revathi, 2016). Liout et al. (1992) proved that CALL assisted introduction was more helpful to L2 learning than traditional introduction. In their findings, Azevedo and Bernard (1995) indicated that computer-feedbacked group performed better than 79% percent of the control group. Nagata (1997, 2013) claimed that computer instructions were more efficient and intelligent feedback proved useful in addition to that, students understood complex grammatical concepts easier with metalinguistic feedback from computers (Revathi, 2016). Furthermore, a study by Plakans (2001) suggested that materials that were developed by the instructors to give feedback seemed more user-friendly.

From past to present, Computer-assisted feedback is found to be valuable for all proficiency levels. Research on feedback was mostly on written format. Kıymaz (2022) showed an alternative modal to written feedback with audio feedback using research data on audio feedback. However, she found that a considerable amount of feedback may lack effectiveness and efficiency, potentially because of the traditional approach that emphasizes providing feedback in written form. (Kıymaz, 2022; McCarthy, 2020) She claimed that written feedback cannot match the needs of 21st century students who are “digital natives”. Dixon (2015) explained audio feedback as a form of feedback where the instructors record their voices and send them to students’ tasks. Although there is no clear consensus on

the usefulness of audio feedback, it is an important area to figure out in CALL feedback scenarios.

There are numerous ways to use CALL technologies to help teachers' feedback tasks. However, the nature of computer systems thus far supports written feedback more than the other types of feedback. One of the reasons for researching favouring written feedback in CALL that others is doubtlessly *Automated Writing Evaluation* (AWE) feedback system.

1.5.11 AI and Feedback

“Developments in technology, particularly in the field of natural language processing (NLP) and latent semantic analysis (LSA), have led to the increased uptake of software that can automatically analyse student writing” (Hockly, 2019, p. 1). Language education fields have adopted the computer technologies to benefit from them in many aspects and NLP has been used for both holistic scoring and writing evaluation (Feng et al., 2015; Parra & Calero, 2019). As Warschauer and Grimes (2010) pointed out, one of the most useful tools to serve the purpose of effective assessment of writing is automated writing evaluation (AWE).

Researchers have seen a potential in automated writing systems, and it has been an increasingly popular field of EFL, ESL and ELT (Chen & Cheng, 2008; Warschauer & Ware, 2006; Yang, 2004; Cotos, 2011; Hockly, 2019, Parra & Calero, 2019). One of the most used functions of the connection between AWE and L2 writing is automated feedback, which was thought to be a complementary tool to teacher's feedback initially (Cotos, 2011; Shermis & Burstein, 2003; Ware, 2005; Warschauer & Ware, 2006). AWE programs have the ability to deliver instant feedback both qualitative and quantitative, they can successfully grade the text, analyse the text mostly in structural form, and create a detailed evaluation (Khoii & Doroudian, 2014; Parra & Calero, 2019). AWE systems use large databases to compare written texts, the database includes thousands of written texts on the same genre of different levels and using correct prompts allows the system to work within to effectively analyse the given text with thousands of others, cross-referencing and comparing them. Additionally, using statistical modelling and NLP algorithms, AWE systems analyse the syntax, coherence, word count and even vocabulary range of the texts then provide an overall grade, suggestions and even feedback. Eventually resulting in a complete analysis of the given text (Hockly, 2019).

One important advantage of AWE technologies is that they help teachers with their labour significantly reducing the workload, some other advantages are that AWE technologies have consistent explanations and immediate feedback opportunities helping students to increase their autonomy by giving them the chance to control the pace of their own learning (Bai & Hu, 2017a; Hyland & Hyland, 2006; Liao, 2016a; Parra & Calero, 2019; Warschauer, 2000; Warschauer & Ware, 2006; Zhang & Hyland, 2018) Also, the consistent and objective results help learners to improve their mechanical use of language and accuracy in writing (Parra & Calero, 2019; Y. -J. Wang et al., 2013).

One of the main reasons why researchers and teachers get assist from computer technologies is that teacher feedback or peer feedback is not consistent all the time due to the human factor, human error may be a detrimental factor for students and they might find themselves confused (Lavolette et al., 2015; Ranalli et al., 2017; Parra & Caleroa, 2019; Zhang, 2016). Burstein et al. (2004) explains that the best way to improve writing is recurrent practice and accurate, appropriate, consistent and immediate feedback. Regarding all, AWE systems provide a great support to teachers and flexibility and better time management opportunities to students (Shim, 2013). Because of all the advantages that AWE delivers, its value as an assistant that is able to ensure high-level feedback and significantly improve writing quality, has been increasing (Wilson & Czik, 2016).

Y. -J. Wang et al. (2013) explains the importance of AWE by attributing AWE as an essential computer system that helps learners to improve their writing because of its quick and personal feedback chances with detailed explanations and corrections of grammar, spelling, sentence structure and vocabulary choice in detail eventually helping students' autonomy. Collection and analysis of some studies indicate three major advantages of using AWE, these are: (1) First, word processors that aids in editing and revising of basically mechanical use of language such as spelling and grammar, helps students to increase their self-awareness of what they write. (2) Second, error correction which enables students to review their errors instantly and gives teachers the opportunity to communicate with students directly for their errors and feedback. (3) Last, computer technologies and artificial intelligence systems are arguably a lot more objective and accurate especially when grading essays.

Although the human feedback was seen as more flexible and better addressing the students' backgrounds and needs recent technologies are improved to a point where they can ensure individualistic feedback more effective. Furthermore, these systems can check many numbers of essays almost immediately and grade them effectively benefiting from latent semantic analysis (Khoii & Doroudian, 2013; Shim, 2013; P. Wang, 2013; Parra & Calero, 2019; Wang & Wang, 2015; Warschauer & Grimes, 2010). Although there are some concerns about the validity and effectiveness of AWE systems general research on the systems is fairly positive. Additionally, the latest instalments and versions of these system, especially after the drastic progress of AI technologies with trendy chatbots, are getting closer to human-like conversations closing the gap between the previous concerns and the latest results.

AWE programs have been used for both summative assessment and formative assessment purposes (Hockly, 2019). AWE has a rather debated spot in summative assessment studies. As aforementioned, benefits of AWE are similar in summative assessment as well. It is extremely faster and cheaper thus it can assess a large number of high-stake tests in little amount of time. Moreover, studies show that despite some exceptions human and machine scores are fairly similar in grading essays (Elliot et al., 2013; Hockly, 2019). However, there are some questions about the capabilities of AWE systems in terms of whether they can evaluate "...critical thinking, rhetorical knowledge, creativity, or a student's ability to tailor their text to a specific readership." (Hockly, 2019) Herrington and Moran (2012) showed that one of the systems performed poorly whilst scoring their essay accurately.

On the other hand, formative assessment studies favour AWE technologies especially in ELT to support students' writing skills (Hockly, 2019). Studies indicate that the practise of AWE programs for formative assessment can boost students to revise their work and their motivation to write (Chapelle, 2001; Warschauer & Grimes, 2010). J. Li et al. (2015) discovered that *Criterion*, a writing evaluation model, led students to revise their written work significantly more, and due to its corrective feedback, help students to improve their accuracy over drafts (Hockly, 2019). Another study by Parra & Calero (2019) found that the use of AWE programs, in this case *Grammarly* and *Grammarly* significantly helped students to improve their writing proficiency. They made less errors and their awareness of the written work increased. Furthermore, the feedback provided by the used

programmes aided the students to enhance their "...grammar, punctuation, mechanics, and style, and therefore, their writing accuracy." (Parra & Calero, 2019). In addition to that, students had positive opinions towards using AWE technologies as well as teachers. Regarding the general results of their study Parra & Calero (2019, p. 222) summarises as such:

In conclusion, according to the results of this research, the AWE programs Grammarly and Grammarly enhanced significant aspects of the writing skill on the students' writing performance, motivating them to reformulate their errors by the sense of awareness and autonomy that the learners are conscious when using these tools without the support of the teacher's feedback. However, it is worthy to say that human guidance is required to compensate the limitations of AWE programs have, making it clear that these technological tools are classroom aids that complement the writers' improvement. In addition, it is important to understand that writing motivation and writing quality are characteristics that are linked with the objectives and learners needs. Although, AWE tools have demonstrated to have its limitations such as content development, these tools may have benefits for both teachers and students with appropriate teacher's monitoring and guidance.

General view on the use of AWE programs is mostly positive. AWE tools help learners to improve their language learning process in writing, giving them the autonomy of their own learning progress (Wang & Wang, 2015). They explain the given feedback in detail and their feedback is relevant (Dikli & Bleyle, 2014). AWE programs appear to be helpful in positively affecting students' writing quality (Y. - J. Wang et al., 2013) In addition to that, AWE tools eased the teachers' workload and saved time by providing almost immediate evaluation, which could be a huge issue for teachers to give relevant, constant, and accurate feedback in short amounts of time (Chapelle et al., 2015; Parra & Calero, 2019; Warschauer & Grimes, 2010; Wilson & Czik, 2016;). Nevertheless, despite its many advantages there are some doubts about the AWE programs. Especially the earlier forms of AWE programs sometimes lacked authenticity and context. They sometimes favoured a certain type of English which could cause problems for EFL learners and the field of ELT (Hockly, 2019). The programs struggled to give meaningful responses to the rhetoric aspect of the language. However, recent improvements in AI technologies are slowly taking over the problems that earlier programs had. This will eventually lead AI technologies to become more prominent and the use of AWE programs with enhanced AI to provide feedback will become more and more accurate and on-

point, to a point that human feedback will ultimately fall behind. Without a doubt, teachers of ELT and learners will benefit from the oncoming technologies.

Speaking of upcoming technologies, obviously artificial intelligence is the biggest trend in the world at the moment, and it seems like AI will remain as a centre of attention for many more years to come. Of course, AI is nothing new as aforementioned, moreover many of the AWE tools benefit from AI technologies as well. One of the greatest geniuses of the last century Alan Turing proposed a question in the 1950s, “Can machines think?” (Turing, 1950 p.433) This question is yet to be answered however humans rely on AI for many purposes. AI’s opportunities aid educators, teachers, researchers in many areas of education and as an essential part of the education system, feedback is one of them. Especially after the publication of OpenAI’s chatbot, ChatGPT, which is an interactive AI chatbot that uses GPT-3,5, GPT-4, and recently GPT-4o language models that is able to effectively, accurately, and most importantly humanely respond to the many requests of users such as having conversations, translating texts, generating texts, summarizing information, gathering information, developing computer programs, creating digital art, giving feedback, and many more (Cotton et al., 2023; Gao et al., 2023; Khalil & Er, 2023; Lund & Wang, 2023; Ibrahim, 2023). The most remarkable feature of ChatGPT is that its ability to produce coherent, intelligible, human-like language and complete cognitive tasks (Cotton et al., 2023; Susnjak, 2022). After OpenAI, in short amount of time many companies have published their own version of chatbots; BardAI, NotionAI, ZacAI are just a few of them. What differs AI from normal computer programs is that computer programs are “programmed” to effectively complete a task, for instance, giving feedback. Computer programs can successfully give feedback, evaluate, and analyse written sources almost with exact results every time. On the other hand, AI technologies constantly improve their responses due to their complex algorithms improving their ability to provide more effective and accurate feedback every other time. Generally, AI-enhanced tools are programmed to do one specific task, in this case, giving feedback. The difference between AI-enhanced tools and artificially intelligent chatbots is that, despite AI-enhanced tools, recent artificial intelligence technologies and as a sub-field of those technologies, chatbots can do almost anything. In the context of feedback ChatGPT provides AI-generated feedback (Farrokhnia et al., 2023). Although the literature suggests that AI-generated

feedback certainly holds promise and has immense potential there is still not enough evidence to effectively support this claim (Dai et al., 2023). However, the research on AI-generated feedback is showing incredible growth and will show exponential growth in the very near future.

Banihashem et al. (2024), conducts a study on the effectiveness of ChatGPT as a feedback tool for students, comparing it to peer feedback. Banihashem et al. (2024) explains that the potential of ChatGPT to provide quality feedback could supplement the feedback given by peers, thereby improving feedback methods in higher education. This is especially useful for large classes with significant essay-writing elements, as it gives educators a practical option to deliver valuable feedback to a greater number of students (Banihashem et al., 2024). Additionally, Mizumoto and Eguichi (2023) discovered that ChatGPT showed huge potential as an automated essay scoring tool. They further claimed that scoring and feedback by ChatGPT "...can be considered as a viable alternative for evaluating and providing feedback on L2 writing." (Mizumoto & Eguichi, 2023, p. 11) Another study by Steiss et al. (2024) found that although well-trained evaluators delivered better feedback with more sophisticated quality, they proposed that the overall quality of the given feedback and the simplicity of the process would still be useful in some instances such as early drafts or when there is no well-trained evaluator. However, it is worth mentioning that instructors especially in higher education struggle to provide high quality feedback consistently mostly because of time limitations (Boud & Molloy, 2013; Dai et al., 2023). Another study by Dai et al. (2023) supports this claim by showing that ChatGPT is able to "...generate more readable feedback with better consistency, which helps teachers to provide personalized feedback with higher quality for larger classes in less time." (Dai et al., 2023). Also, it is worth noting that Soloyev (2024) aimed to compare students' preferences on the feedback they got from peers, ChatGPT, and teachers. Accordingly, students preferred feedback from ChatGPT and teachers WCF over peer feedback either teacher only or teacher and ChatGPT combined.

The field of AI-generated feedback is still fresh, but studies show promise that it can be useful for students and teachers. Since AI models are constantly changing and improving themselves, with more data it is possible to achieve better outcomes in the near future, and perhaps almost excellent results in not so far future.

2. MATERIALS and METHOD

The aim of this section is to provide the necessary details of the methodology used in the research. It provides an overview of the research design, study's participants, the tools and steps used for data collection and data analysis. Furthermore, it explains the trustworthiness of the study and the role of the researcher in the process of the research.

2.1 Research Design

Mixed-method research have the potential to uncover significantly more reliable and more precise data compared to single method studies (Creswell, 2004). Mixed method allows researchers to inspect the data both qualitatively and quantitatively eventually resulting in sound results. Furthermore, mixed-method research can answer multiple questions related to the study. "The complexity of the modern world demands more sophisticated approaches to understand it." (Ivankova & Creswell, 2009, p. 136) successfully explains why researchers may need this strategy in their studies. In addition, they elucidate as such:

Mixed methods research, with its focus on the meaningful integration of both quantitative and qualitative data, can provide a depth and breadth that a single approach may lack by itself. Combining two methods would also increase the reliability of the results significantly. However, because of this integration, mixed methods research can often prove challenging for novice researchers.

Although there is a general view of qualitative data being hermeneutical and quantitative data being positivist, Onweuegbuzie and Leech (2005) opposes this view. Cohen et al. (2018) argue that "...quantitative approaches can catch opinions, perceptions, probabilistic causality and process approaches, and qualitative approaches can feature in experiments, identifying causality, surveys and patterns of, and trends in, data." However, it would not be totally wrong to generalise quantitative data as the numeric data that provides valuable information about the patterns of a research, and qualitative data as the tool to perceive opinions and perceptions of groups of subjects.

As for this study, quantitative data was used to determine the patterns of teachers and AI-based chatbots in grading essays. The quantitative aspect of the research provided invaluable information on recognizing the similarities and differences of grading even while using the same tool to grade. On the other hand,

qualitative data was used to get a deeper understanding on the perspectives via given feedback by teachers and chatbots. Moreover, the reasons behind the grading and feedback were further explored to better respond to the “why” questions.

Consequently, it is obvious that feedback in ELT is better explained with mixed-method strategies. The researcher benefited from this strategy in order to obtain valuable insight into teachers’ perspectives, envisioning getting a sound grasp on understanding the differences between human feedback and AI-based feedback. Strategies used in this research particularly aims to answer following research questions:

1. How do humans grade and give feedback to different types of essays?
2. How do AI-based chatbots grade and give feedback to different types of essays?
3. What are the differences and/or similarities between human grading and feedback vs. AI-based grading and feedback on written essays?
4. Are there any differences among AI tools – ChatGPT, BingAI, and Gemini- in grading and feedback?

2.2 Participants

This study was conducted with two different groups. The first group has 5 English language teachers who have been working for 2 to 5 years who were asked to evaluate 30 essays. The teachers vary in age from 24-30, and all have teaching experiences. They are familiar with the essay evaluation process and using rubrics due to formal education in the last year of education faculty that was taken by all teachers. All the participants voluntarily accepted to be a part of this study.

2.2.1 Tools

The second group is recently developed Artificially Intelligent chatbots. Currently a large variety of chatbots are being used, however, for this study, ChatGPT 3,5 developed by OpenAI, BingAI, which is an extension of Bing enhanced by ChatGPT 4,0 modal, and Gemini created by Google were used. All the programs are free to use and highly accessible.

1. **ChatGPT:** ChatGPT is an advanced conversational AI model developed by OpenAI, first launched in November 2022. It uses a neural network architecture called GPT (Generative Pre-trained Transformer) to generate text based on context and prompts. Trained on a vast dataset encompassing diverse topics, ChatGPT is designed to simulate human-

like conversation and provide detailed, contextually appropriate responses. Its evolution from earlier GPT models allows it to handle increasingly complex language tasks, including content generation, language translation, and writing evaluation. ChatGPT excels in providing personalized, coherent, and creative writing feedback, focusing on grammar, organization, and overall readability. It can assess writing based on pre-defined criteria and assign scores.

2. **BingAI:** Bing AI is Microsoft's AI-powered assistant, officially introduced in February 2023 as part of the enhanced Bing search engine. It combines advanced natural language modeling with real-time access to web search data, enabling it to provide responses enriched with up-to-date information and contextual understanding. By leveraging its connection to the web, Bing AI offers highly relevant suggestions for writing tasks, aligning its advice with the latest standards and conventions. Furthermore, it is seamlessly integrated with tools such as Word and Teams which makes it easily accessible for educators and students. It provides structured, real-time feedback emphasizing clarity and alignment with modern writing expectations.
3. **Gemini:** Gemini AI, developed by Google, was unveiled in December 2023 as a next-generation AI system. It integrates large language models with real-time data retrieval and multimodal capabilities, combining text, images, and videos to offer versatile and enriched feedback. Designed to provide nuanced, context-aware responses, Gemini draws from Google's extensive ecosystem and live search capabilities, ensuring its suggestions are both accurate and timely. This makes it particularly effective in addressing complex or multidisciplinary writing tasks. Gemini is able to incorporate live, fresh data from Google Search which enhances the relevance and specificity of its feedback. Its support for a wide range of languages and multimodal inputs makes it a comprehensive tool for diverse academic and professional needs.

The goal of this mixed-method study is to have a common understanding of the capabilities of recent AI technologies by comparing results, grading, and feedback of 30 essays that are 10 narrative essays, 10 expository essays and 10

argumentative essays. Although there are strict procedures to follow whilst grading, the researcher cannot possibly claim that the results would be similar in any study.

A convenience sampling method was used to select the first group of participants. One of the main reasons to choose this technique is because of the ease of access to participants. This method involves selecting participants who are accessible and willing to participate. The second group that consists of chatbots were chosen via purposive sampling. The chatbots were selected due to being free, accessible, and the most popular and arguably the best examples of all chatbots.

2.3 Data Collection

This research includes two different types of datasets. The first set is the essays which was collected from the students, and the other set is the evaluation of those essays which was collected from the teachers and artificial intelligence tools.

2.3.1 Essay Writing

This dataset consisted of 30 essays in total written by 10 first-year ELT students from Bolu Abant Izzet Baysal University who volunteered to participate in this research. Each student was required to write three types of essays:

1. **Narrative Essays:** The main purpose of narrative essays is to tell stories or write about personal experiences. Narrative essays mostly focus on using imagination and creativity in writing. Narrative essays are rather flexible in terms of structure and the use of language. In this context, narrative essays are particularly relevant because they offer a rich medium for analyzing writing skills such as coherence, creativity, and lexical range. The open-ended nature of narrative essays allows for diverse interpretations, making them an ideal choice for evaluating feedback provided by AI and teachers on both the technical and creative aspects of writing.
2. **Expository Essays:** Expository essays aim to explain a topic with a neutral approach. They are the expressions of knowledge or expertise in a specific area. There are no emotions or personal opinions in expository essays rather the focus is on facts and evidence. Expository essays have fairly rigid structures and formal language. This study benefits from expository essays due to their relevancy for evaluating analytical and organizational skills in writing. The structured format provides clear criteria for assessing aspects such as coherence, factual accuracy, and

language precision, which can be considered essential when comparing the feedback provided by AI tools and teachers.

3. **Argumentative Essays:** Argumentative essays are one of the most common types of essays. The goal in argumentative essays is to convince the reader to the point of view of the reader using well-structured argumentation and evidence-based claims. The essays should include research about the certain topic, arguments and counterarguments, and evidence. The writers should avoid expressing emotions and use proof to support their opinions. The language should be non-colloquial and there should be a clear and strict structure. In the context of this study, argumentative essays are particularly useful for examining critical thinking, logical organization, and the effective use of evidence in writing. Their emphasis on structure and rigor provides a clear framework for comparing the feedback offered by AI tools and human teachers, highlighting how each addresses aspects such as the coherence of ideas, the validity of arguments, and the appropriateness of language.

Choice of the topics were left to the students to collect more personal and original essays. However, there was a constant consultation between the researcher and the students to aid them in finding relevant and appropriate topics.

2.3.2 Essay Evaluation

The evaluation process involved two main stages: human grading and AI-based grading.

1. **Human Grading:**
 - a. Five English language teachers were asked to grade the essays. The teachers had prior knowledge and experience in teaching and evaluating writing. The experience of the teachers differed.
 - b. The teachers are presented the TOEFL IBT writing rubric to evaluate the essays. The rubric focuses on four main criteria (ETS, 2020).
 - i. The first criterion is the task response which is the essay's ability to effectively address the required topic within the

essay and paragraphs without deviating from the main subject so much.

- ii. The second criterion is organisation that refers to the structure of the essay. Organising ideas clearly and logically is the critical goal of this step
 - iii. The third criterion is coherence and cohesion factors evaluating the unity in an essay. Cohesion is the result of the smooth flow of ideas, sentences and paragraphs of the essay.
 - iv. The last criterion is the mechanics of the language which refers to the grammar, vocabulary, and pragmatics aspects of the language used in essays. Diversity of the grammar rules, appropriate and correct wording, syntactic variety, use of idioms and flexible language are focused to evaluate mechanics.
- c. Each essay was independently graded by the teachers to ensure inter-rater reliability. The scores were then averaged to obtain a final grade for each essay. Additionally, the feedback for each criterion and the overall feedback were inspected to analyse the issues of the essays that was focused by each teacher to get a wholistic understanding of the essays from different perspectives. Using multiple raters and averaging scores is a common practice to enhance reliability (Jonsson & Svingby, 2007).

2. AI-Based Grading:

- a. The essays were graded using three different chatbots; ChatGPT-3.5, BingAI, Gemini. The results were examined and cross-referenced to find out the common points of AI-generated feedback.
- b. Chatbots were instructed to grade the essays according to the TOEFL IBT writing rubric. The same criteria used by human teachers were employed to maintain consistency. The instructions to chatbots included detailed explanations of the rubric criteria to ensure accurate and relevant grading and feedback.

Both human grading and AI-Based grading utilised the TOEFL IBT writing evaluation rubric. TOEFL IBT's rubric is highly recommended, and it has high

reliability and validity (Plakans et al., 2019). It grades four criteria; task response, organisation, coherence and cohesion, mechanics of the language, from 0 to 5; 0 being the least and 5 being the most responsive to the tasks and criteria.

2.3.3 Data Collection Procedures

As aforementioned, this study has two types of datasets. The first set is the essays written by ten first-grade students from Bolu Abant Izzet Baysal University and the second set is the grades and feedback from ten teachers and AI enhanced chatbots and teachers' personal comments and opinions about their grading and feedback.

To begin with, collection of essays had several steps. First step was to select the group of first-grade students that would voluntarily participate to this research. The way students were selected was convenience sampling method. The researcher asked all of the first-grade students and selected the first ten volunteers. The researcher opened a WhatsApp group to communicate with the participants. After the WhatsApp group, the researcher prepared a short video and a presentation using Canva on the processes of writing an essay. Other step was to prepare videos and presentations for each essay type. Short videos and presentations were sent to participants via WhatsApp group. Videos and presentations were sent week by week for three weeks. The first week was for narrative essays, and the participants were given one week to complete and send the essays to the researcher. The second week was dedicated to expository essays, and the last week was for argumentative essays. Videos and presentations were sent in the first day of the week and an online meeting was conducted on the second or third day of the week when necessary. During the week, the participants were asked to send their tasks via WhatsApp or e-mail. Most students used WhatsApp to send their tasks. All of the videos and presentations were uploaded into an online drive for ease of access. Lastly, all the essays were uploaded into another drive folder for other purposes. With that, the collection process of essays was completed.

Second process was to collect the evaluation data. In order to do that, crucial step was to find teachers. Teachers were collected using convenience sampling method, the researcher asked different groups of teachers to find volunteers and volunteers were informed of the process with personal messaging tools. An online classroom platform was created using Google Classroom. The teachers were asked

to join this classroom using a link sent by the researcher. Grading and feedback of the essays were collected simultaneously. The researcher created a word document for each of the essay consisting of both grading and feedback (See Appendix). TOEFL IBT writing rubric was taken as the primary and guiding source for the process. The document and the essays were collected in a google drive folder. The online classroom and the drive folder were linked, and the researcher opened three tasks to collect the evaluation data. The researcher again created a video and a presentation explaining to process of writing evaluation and TOEFL IBT writing evaluation rubric. The teachers used the tasks personally opened for them to evaluate the essays. Afterwards, semi-structured interviews conducted with the teachers to better understand the process of their evaluation and get a greater grasp of the personal views of teachers on evaluation, grading and giving feedback. These interviews were recorded and conducted using Zoom then they were scripted afterwards.

Lastly, the other part of evaluation data was collected with AI-enhanced chatbots. Chatbots were selected due to their popularity and success. The most famous and successful chatbot perceived during the time was ChatGPT. In order to get more data on the evaluation of AI-enhanced feedback additional chatbots were used. These chatbots were other popular chatbots besides ChatGPT that are BingAI and Gemini. Prompt engineering is one of the most critical processes for getting better results via chatbots. For the consistency of the results, the same prompt was used for all chatbots. To enhance the accessibility simple prompts were used without detailed and specialized prompt engineering. The prompt consisted of two parts. In the first part, TOEFL IBT writing evaluation rubric was introduced to the chatbots, and the second part demanded chatbots the evaluate given essays according to the rubric. Below is an example of a prompt for the evaluation of a narrative essay.

“Based on the TOEFL IBT Writing Evaluation Rubric give feedback to and grade the given narrative essay. Consider that the student is a first-grade ELT student at Bolu Abant Izzet Baysal University.”

This specific prompt was used for all essays only adjusted for essay types for all chatbots. After the prompt, the essay was given in text form then the results of the grading and feedback were collected. Lastly, no additional commentaries and

tinkering were made to get optimal results, all answers were taken directly in their first forms.

The collected data were categorised and uploaded into different folders. Data collection process was completed after the essays, grading/feedback and interviews were completed. Thus, the end of data collection was completed successfully.

2.4 Data Analysis

This section will provide the details about the analysing process of the data that was collected using mixed-method strategies.

2.4.1 Quantitative Data Analysis

This study benefited from quantitative data to have an average understanding of the grading results and to support given feedback. Quantitative data analysis was conducted on the grading of the essays. Grades that were given using the TOEFL IBT writing evaluation rubric by teachers and chatbots were separated and categorised. The reason of this separation was to get the average points that was granted by teachers and chatbots. After the separation, the grades were then analysed using the Statistical Package for Social Sciences (SPSS) to find the mean grades for each criterion on the rubric for both teachers and chatbots. Descriptive statistics, mainly mean scores were used in the analysis of the quantitative data.

2.4.2 Qualitative Data Analysis

Qualitative data analysis was conducted according to the given feedback by teachers and chatbots that were assigned using the TOEFL IBT writing evaluation rubric under the given criteria. In addition to the feedback on criteria, overall feedback was given both by AI and teachers. Lastly, for further information an interview was conducted with teachers regarding their additional comments. These commentaries focused on the evaluation process and the steps that teachers took during the process.

The findings from the quantitative and qualitative analysis were merged in order to find any possible differences between teachers and AI-enhanced chatbots in terms of grading and feedback and to shed a light on the possible reasons of these differences and/or similarities.

2.5 The Trustworthiness of the Study

This study aims to investigate and identify the differences and/or similarities between teachers and AI-enhanced chatbots in terms of grading and giving feedback on three different types of essays written by first-grade ELT students. In order to ensure a better understanding, feedback was asked besides only grading. Furthermore, for a deeper insight of the supposed differences additional commentaries by teachers were given. The analysis of the data meant to enlighten the dissimilarities allowing sounder comprehension on the issue.

2.5.1 The Role of the Researcher

The researcher acted as an enforcer in this project. The main purpose of the researcher was to guide both the students and the teachers when needed. In order to ensure reliability, the researcher avoided certain interruptions. Furthermore, the researcher did not comment on the essay writing process and the essay evaluation process aiming not to affect the teachers and the students in any way. Objectivity was affirmed by providing only presentations and videos and further explanations on the writing process and the evaluation process of essays.

3. RESULTS

This part will be devoted to explain the results of the study by exploring the grading and feedback made by human teachers then by AI enhanced chatbots. Analyses of the gradings and feedback will be compared to find possible patterns, similarities, and differences between AI and human teachers and then will be explained in detail.

3.1 Analysis of Human Grading

In this section, human grading will be explained as the thesis's quantitative data. First, the results of the narrative essays will be explored, followed by the expository, and the last essay type will be the argumentative essay. The process of grading for all essay types is the same. For both grading and feedback, TOEFL IBT Writing rubric was used. This rubric considers four main criteria: Task Response, Organisation, Coherence and Language Use in addition to these four criteria there is also a part for Overall Feedback was added. For each part, graders were asked to grade the criterion in points between 1 to 5. In this regard, 1 or lower grades mean the essay completely failed to meet the needs of the graded criterion and 5 means that the essay successfully managed to address the necessities of the graded criterion. As for the last, all the points that were given for each criterion then added and a point out of 20 is given for the final score in overall feedback. For instance, if a student gets 3 points for task response, 4 for organisation, 2 for coherence and 3 for language use, total point of the student is 12/20. This total number is written in the overall feedback part.

3.1.1 Narrative Essays

Amongst the types of essays, narrative essays are slightly different due to its flexible nature as an essay type. Generally, essays follow a precise structure and procedure with the use of more complex and non-colloquial language. However, narrative essays aim to narrate and tell a story as vividly as possible. Consequently, narrative essays may benefit from more colloquial language and flexible structure. Furthermore, narrative essays are told personal experiences or memories that is why they are reflections of writers. As a result of that, grading and giving feedback to narrative essays may slightly differ than other types of essays where structure is heavily regarded.

Furthermore, in an interview with one of the teachers; the teacher mentioned that the narrative essays were the most difficult to grade due to them being a reflection of students. She added that:

“Students may feel a word that they used is the correct word for that place, when I read it sometimes feel awkward, but since I am not totally aware of the students’ mindset, it was difficult to evaluate personal narratives.”

Although they are indeed personal reflections of the students, TOEFL IBT writing rubric manages to limit essays in a frame successfully. This is a crucial factor in evaluating to ensure inter-rater reliability. Results of the analysis show great consistency amongst each essay. Analyses show that there is little standard deviation between grades and overall points are always in similar ranges. However, not all the gradings are the same between teachers. For instance, in an essay that was graded 18 total by two different teachers, one of the teachers gave 4 points to language use and task response whereas the other teacher gave 4 points to organisation and coherence. Needless to say, further explanation required to address the differences, but these will be explained in the “feedback analysis” part of the results.

As for the grading, although there are differences between essays, overall grading is quite consistent and given points are relatively similar to each other except for a few essays. In many of the essays, there are not huge gaps between gradings for each criterion. Although some essays, have slight differences in grading with a standard deviation of being near 1, many of the essays are quite close with a SD of 0,5 or lower. Analysis of the data indicates that the most consistent graded criterion regarding all of the essays is the grading of the task response with a SD of 0.13 and the least consistent is the overall grading with a SD of 0.37 additionally the grading of language use follows overall grading with a similar SD.

In conclusion, human grading in narrative essays reveal that there is a significant consistency of gradings between raters. Each criterion is graded with little difference between teachers. Overall grading follows a similar trend of grades making gradings reliable and indicates that human-graders follow parallel thought processes whilst grading.

3.1.2 Expository Essays

Although expository essays are a lot more similar to argumentative essays than narrative essays, the differences between expository and argumentative essays should be taken into consideration since they aim different goals, distinct purposes, and unique writing styles. Main goals of expository essays are to explicitly describe and explain certain topics in detail with utmost objectivity. In expository essays, authors must not defend an idea, a belief, or an opinion on the contrary they must show a topic clearly presenting clear ideas with a solid order of ideas, strong structure, and great coherence. Grammar and the variety of language mechanics are important, additionally the language should be formal and precise.

Consequently, the evaluation of the expository essays considerably differs when compared to narrative essays. Therefore, even when using the same rubric, evaluators take different aspects into account. To simply put for instance, the task for narrative essays is to successfully narrate an event or a memory with vivid descriptions whereas expository essays want to expound subjects in detail.

Analysis of the expository essays was slightly different. Although the results show that there are many criteria of some essays had 0 SD, overall SD of almost every criterion is more than narrative essays. Lowest SD rates found in Language Use with 0.17 and highest in overall rating with 0.9. One crucial finding is that some expository essays divided graders literally into two, half of the graders for some essays graded above average which is 3, and other half are below average.

All in all, statistics indicate there is a slight difference of thought in grading expository essays. Analysis also strongly reveals that teachers evaluate expository essays similarly, with comparable grades for each criterion and overall point. Exceptions are rare and even for exceptions overall gradings are not dramatically different.

3.1.3 Argumentative Essays

The last type of essay is the argumentative essay. As aforementioned argumentative essays are similar to expository essays in terms of structural features. Argumentative essays too, follow strict rules in structure. There must be clear organisation of ideas with defined introduction, body, and conclusion paragraphs. Use of language is similar to expository essays with a certain need for variety of grammatical and lexical items. Formal language is important to use in argumentative essays.

However, difference begins with the handling of the topics. It was stated that expository essays should be objective and present a topic, an idea, or a belief without clearly expressing the authors' ideas, on the other hand argumentative essays handle the topics more differently. The main purpose of argumentative essays is to form argumentation defending a certain belief, opinion or a position. In argumentative essays it is crucial to create argument-based essays. Arguments should be supported by evidence and further data. Additionally, counter-arguments need to be mentioned in the essay as well. Writers want to persuade readers of their position and their view about a certain subject that is controversial.

Given all the differences, it is natural to expect different thought processes whilst evaluating argumentative essays. Consequently, analysis of the data reveals this difference. According to the analysis, argumentative essays are the essays with the biggest variety of gradings. As a result of this, standard deviation for each criterion is slightly higher than both of the essay types. Similarly to expository essays, the least difference is found in the grading of language use with a SD of 0.2 also the highest difference is within the overall grading, moreover overall gradings of the argumentative essays have the highest standard deviation in all of the analysis including expository and narrative essays, with a 1.5 possible deviation.

Analysis of the data indicates that teachers have the least amount of compatibility in the grading of argumentative essays. Accordingly, the compatibility is lower than other essay types. However, it can be claimed that most of the teachers follow similar procedures whilst evaluating and grading argumentative essays.

3.1.4 Conclusion

All things considered, analysis of the data reveals that amongst teachers of this particular group, narrative essays are the most compatible types of essays and argumentative essays are the least compatible types between narrative essays, expository essays, and argumentative essays. Data further indicates that although some essays are controversial in terms of grading and have different views, it can be seen that consistency between teachers are relatively high. However, certain essays especially in argumentative essays, show great inconsistency between the graders. On the positive, number of those essays are covering a small percentage of total essays. In conclusion, this specific group of teachers show a relatively nice range of reliability and consistency within three given essay types and thirty essays

in total. Although it is still worth mentioning that not all essays have the same reliability.

3.2 Analysis of Human Feedback

This section will be devoted to explaining the qualitative data of the study that is the given written feedback of the essays. First of all, the feedback to narrative essays will be shared followed by the feedback to expository and argumentative essays. Like abovementioned, the feedback was given taking TOEFL IBT writing evaluation rubric as basis. Four criteria were first graded then feedback for each criterion was given by evaluators. Then the evaluators gave overall feedback to writers to further explore essays. Optimally, given feedback are consistent with the grading and will be of aid to detailly explore the positives and negatives of each essay in detail.

3.2.1 Narrative Essays

It is without a doubt that feedback is more personal, detailed, and specific than just simply grading. Furthermore, the way of giving feedback changes depending on the essay type. Evaluators focus on different aspects of different types of essays moreover they consider distinctive features when evaluating. For instance, narrative essays require a more story-driven coherence whereas argumentative essays expect a more structured and defined coherence.

In this regard, it is expected to find different focal points, keywords, and responses to the same criteria. Analysis of the narrative essays indicates that human evaluators deem structural aspects of narrative essays as the most important aspect of a narrative essay with over forty mentioning. Structure in this regard means all organisational aspects of an essay from paragraphing to general design of the essay. Following structure, requirements of a narrative essay is the second most criticised point with exactly forty feedback highlighting the importance of it. Requirements is an umbrella term that covers mostly the task response of the narrative essay. It evaluates how well the essay represents the qualifications of a narrative essay. Syntactic and lexical variety and grammar rules involved in a great number of feedback and considered vital in narrative essays. More specifically to those mentioned above, the feedback in the task response part generally criticises requirements of the essay whether it presents an effective and successful example of a narrative essay or not. Organisation part mainly focuses on the structure in addition to ordering. Chronological and logical ordering is mentioned to be crucial

in organisation of a narrative essay. Dividing essay to different paragraphs and finding the correct sequence is also considered vital. In coherence part, evaluators emphasis on the flow, harmony and unity of the essay. As for the language use, grammar, spelling and word choice are mentioned but feedback does not have many examples of the wrong uses. Lastly, overall feedback analysis indicates that evaluators mention the biggest missing parts of the essay that were aforementioned in other criteria. If the essay seriously lacks in unity or tone, overall feedback further explains those missing or “worse” points. However, for some essays evaluators reflect on the positive sides of the essay as well.

Additionally, analysis of the narrative essay feedback shows that out of all the comments and feedback there are 107 positive comments in total 91 negative comments and only 15 suggestions were made. The analysis did not reveal any specific reason for this. Some essays which were graded 15 or more received negative feedback as for some essays with lower grading had some positive comments. As for the overall feedback, with some feedback having both positive and negative comments there are 29 positive and 25 negative comments in total. There is no clear evidence to indicate the reasoning behind giving positive or negative feedback.

3.2.2 Expository Essays

The analysis of the expository essay indicates rather similar results to narrative essays in terms of positive and negative commentaries. Evaluators gave 100 positive feedback in total and 91 negative feedback in addition to 22 suggestions. However, there are many different opinions in contrast to narrative essays. There are no major differences between different themes. For instance, transition, unity, grammar, connection, and some other concepts were mentioned similar number of times changing between 15-25 grammar being the most mentioned in between them. Yet there is one part with slightly more remark which is structure. Structure is the only part that surpasses 30 references. Evaluators criticised essays on their introduction, body, and conclusion paragraphs and the overall organisation of those paragraphs. Another interesting result is that evaluators took elaboration into consideration and there are 12 criticisms on elaboration.

In conclusion, analysis of the evaluation of the expository essays showed similarities to narrative essays when giving positive, negative comments and

providing suggestions to students. Conversely, expository essays had more themes and more specific points criticised by evaluators indicating that there are significantly more varied and different aspects of expository essays being considered by evaluators.

3.2.3 Argumentative Essays

Argumentative essay evaluations revealed slightly different results of analysis. With 123 positive comments argumentative essays have the biggest rate of positive feedback. Negative feedback and suggestions are similar in terms of number with 90 negative comments and 10 suggestions. One significantly different result is the rise of themes for argumentative essays. The analysis shows a variety of themes, and their frequencies are a lot higher than expository essays. The most important aspect of an argumentative essay can be seen as the requirements of the essay. Evaluators highly focused on the essays' ability to reflect basic principles of an argumentative essay with exactly forty references to the requirements. Another significant feature seems to be the correct use of language that is grammar, syntax, punctuation, and spelling being referenced 34 times by the evaluators. Flow of the essay follows the grammar with a variety of subthemes such as unity, transition, and connection. Last frequent theme can be found converging on paragraphing with structure and format being highly criticised by the evaluators.

Briefly, analysis of the argumentative essays holds a middle ground between narrative essays and expository essays. It has multiple themes and sub-themes in addition to evaluators commonly referencing to those themes frequently. Hence the result indicates that evaluators share similar notions and think alike in terms of argumentative essays.

3.2.4 Conclusion

In conclusion, the analysis of the human feedback reveals that teachers tend to provide more positive comments on the essays regardless of the point they give. Negative comments also cover almost half of the total feedback. However, the results show that the teachers avoid making suggestions whilst giving feedback.

Additionally, considering all of the themes and essays, teachers seem to focus on different aspects of essays depending on the essay type. Although that is the case, the focus on the grammatical aspects of essays remained relatively similar in each essay type. According to the analysis, teachers mostly focused on structural aspects of narrative essays since the most criticisms whether positive or negative

gathered around the structural aspect of the essays, following that, expository essays indicated similar results in which criticisms being mostly about the structure of the essay. However, two structures were slightly different than each other. In narrative essays structural aspects covered a wider range from organisation to paragraphing although in expository essays teachers seem to focus more on just the paragraphs of the essays. Lastly, feedback of the argumentative essays seems to strongly emphasize the essays' ability to meet the requirement of being an "argumentative" essay.

All in all, by looking at the results it can be expected that overall feedback from the teachers tend to be more positive rather than negative. However, it is worth noting that there are too little suggestions. Additionally, further analysis indicates that teachers focus on different aspects of essays depending on few factors the strongest being the type of an essay.

3.3 Analysis of the Interviews

If one thing could be understood from previous analysis that would be the fact that no matter what rubric evaluators use; evaluations from grading to giving feedback are eventually personal commentaries. These commentaries are of course shaped by multiple factors; beliefs, background, experiences, and even a bad day might be considered amongst those factors. Regarding that, it is critical to understand the reasoning behind the comments to enlighten the differences between evaluators in terms of grading and feedback. Peeking into the evaluators' mind would be beneficial to increase the efficiency of feedback consequently would positively impact the learning and teaching process. With this quest at hand, several questions were asked to evaluators concerning the place of writing as a skill, the importance of writing tasks such as essays, the thoughts of the teachers about evaluation and how to approach to essay evaluation, and opinions about the rubric. Additionally, specific questions about the quality of the essays, essay types, essay difficulties, and important criteria were requested. Hence, the analysis of the interviews will provide invaluable data to uncover the questions of "How?" and "Why?" in grading and feedback.

First of all, the place of writing has long been acknowledged by language teachers by considering it as one of the major skills in language learning. When asked about writing, evaluators too recognised the importance of writing as a skill. Although the importance of writing was undeniable, the evaluators had a variety of

opinions about the place of writing as a skill. One evaluator expressed that whilst writing is an important skill when integrated with other skills, it might not have too much value especially in daily lives. She suggested that writing is the most crucial skill in academic contexts where the learner/students need to write constantly. However, the value of writing slightly decreases in normal settings. Yet she further suggested that writing is critical since it is a skill based on the production of language and many sub-skills like vocabulary knowledge and grammar proficiency would positively be affected by practicing writing. Another evaluator implied that the writing significantly improves critical thinking abilities of students and is connected to the speaking skills. She further proposed that "...as you write you speak better and as you speak you write better". Another different opinion was that an evaluator stated that "...writing is a creative action, and it shows how much somebody has learnt something..." Other than that, all evaluators have agreed that writing is an important productive skill that can improve many sub-skills and integrated to the main skills as well.

Obviously, there are many ways to improve this crucial skill. One of the ways in which learners significantly are able to improve their writing skills is without a doubt are essays. Essays are structured and relatively long written products that require specific ways of writing. Different essay types include different approaches to writing. As a consequence, learners must use a variety of skills to complete essay tasks depending on the type of the essay, the subject, and the theme. This puts essays into a special position that separates them from other more simplistic writing types. Evaluators have all agreed that essays are crucial to improve writing skills. The most common reason that was given was that essays follow a structure and a roadmap that promotes organisational skills combined with the ability to use vocabulary and grammar correctly. Another common concept was the importance of correct feedback in essays. Evaluators suggested that feedback can considerably enhance writing motivation and skill when given correctly.

Thus brings the other point which is essay evaluation. All sorts of evaluation are essential when improving a skill. Feedback is amongst the most critical parts of evaluation process and doubtlessly the most specific one. When asked about the importance of evaluations, teachers have suggested similar concepts. Most of the concepts were focused on feedback. According to the teachers, feedback should have a positive tone and be as detailed as possible. They further claimed that

feedback is most effective when teachers personally know of the students and learners. Therefore, they stated that individual feedback brings forth many advantages to students and their abilities. Additionally, when they are asked about their approaches to essay evaluations the answers were quite diverse. One teacher suggested that she would first approach with the question “How would I write if I were given the task?”, she expressed that this process helps her to create a guide to better evaluate the essays, moreover the length of essay regarding the subject would certainly affect her motivation to read the essay. She also stated that when skimming the students’ use of extravagant vocabulary unnecessarily would catch her eye and have a rather negative effect on the evaluation of the essay. Another teacher explained her approach more mathematical or systematic. Her first impressions were revolved around the essay’s structural compatibility and apparent grammatical mistakes. Additionally, a teacher claimed that titles are critical when she begins evaluating essays. She said that she focuses on the title first then quickly skims the essays to find out if the general concepts match the title. Other than these specific factors, all teachers emphasized the importance of the first sentence and the thesis statements of the essays being crucial.

When beginning evaluation, rubrics are essential guides that teachers and evaluators follow throughout the process. In this project TOEFL IBT Writing Evaluation Rubric was used which is highly recommended and has been proven of its trustworthiness multiple times at multiple studies. Teachers were questioned about the rubric’s efficiency for evaluating all of the essays and three different essay types. All teachers seem to agree on the rubric being efficient. They all suggested that rubric was good for evaluating the essays, it was a capable roadmap they followed through their essay evaluation processes. However, they also claimed that the rubric lacked some specific requirements of different essay types. They suggested that though the rubric is great for general evaluation, better rubrics could be used to meet the specifics of the essay types. For instance, a rubric can be created solely to evaluate narrative essays, another for argumentative essays, or one for expository essays. Still, they found the rubric effective and stated that they would use this rubric for their own purposes as well.

Besides general questions, there were multiple specific questions that were asked to teachers. First question was about the quality of the essays. This question meant to understand teachers’ perspectives on the essays whether they were

reflective of the supposed proficiency levels of students or not, moreover it aimed to get a better insight into teachers' mind when they evaluate essays with a certain assumption of a supposed proficiency level in mind. One of the teachers' stated that out of all 30 essays, there was a big difference between the best and the worst written essays. There were a few great examples, and a few bad examples. Considering all that, the overall quality of the essays was quite expected from the first grade ELT students. Other teachers had similar thoughts as well, despite a few distinct examples on the better side; overall quality reflected the students' proficiency levels. One of the teachers provided deeper insight into this question, she stated that in narrative essays the difference was too much between good and bad essays. There were some great examples, and mid-point was only decent. Expository essays had a general problem according to her statement due to "lack of neutrality" and proposed that it seemed like a limit testing for student. Similarly, argumentative essays lacked argumentative quality because of weak supporting ideas and missing evidence. Despite all, common ground of teachers suggests that students' level of proficiency was greatly reflected on their papers, so the essay quality was good.

Another question was to determine which essay types were most difficult to evaluate for teachers and why, and if difficulties had any impact on the evaluation process. Two of the teachers stated that they struggled with narrative essays the most because they included personal views, memories, and had a certain emotional spectrum. These aspects made narrative essays difficult to criticise. Another teacher thought the narrative was the easiest because it did not have strict rules that binds some other types of essays. Two teachers claimed the expository essays were the most difficult to evaluate. One of them connected this to them being scientific and may require additional research another presented the vagueness of the expository essays as a reason for being difficult. Yet for the same reason, one teacher considered expository essays the easiest to evaluate because of its' scientific approach. Another teacher had the least amount of difficulties whilst working with argumentative essays because they were based on argumentations, and she claimed that evaluating certain arguments was considerably easier. Most of the teachers stated that the struggle they had with the essays impacted the motivation and process of evaluation. This question revealed that teachers may struggle in some essay types more than the others significantly impacting the evaluation processes.

Last question aimed to better understand the considerations of teachers when evaluating different essay types. Teachers were asked which criteria they considered more important than the others depending on the essay types. For narrative essays teachers mostly focused on cohesion. When asked about the reason they explained that because narrative essays follow a story or an event, they should have great coherence and seamless transitions, they claimed that this is crucial to not to get lost in the story and to ensure smooth flow of the story. Another present criterion for narrative essays was the language use. One teacher suggested that narrative essays should include vocabulary or a style that would create vivid images in the readers' mind that is why she considers language use as a pivot for narrative essays. Expository essays on the other hand seems to be revolving around the organisation of the essays. Scientific and evidence based neutral nature of the expository essays was a result that teachers considered it as the most important criterion for expository essays. Yet, one teacher deemed language use as the most critical criterion because of the exact same results. She suggested that these reasons require expository essays to be non-colloquial and official pieces of papers. As for the argumentative essays teachers have all agreed on the task response as the most crucial element of argumentative essays. Creating arguments and supporting them, in other words, being argumentative and reflecting the requirements of the essay type were main reasons of choosing task response as the utmost criterion. Besides these answers, one of the teachers thought that task response was the most important criterion for all essay types because of her approach to essay evaluation. According to her the first thing an essay should be is to greatly reflect the requirements of the specific essay type.

To conclude, an interview to better reflect the teachers approaches and thought processes when evaluating an essay was conducted. Although there were some similarities between answers, teachers had distinct ideas and diverse choices of certain questions. Even when the choices were the same, the reasons varied. These questions and answers gave a general idea of teachers' opinions about the essay evaluation and writing.

3.4 Analysis of AI Grading

Artificial intelligence technologies are seen as invaluable assets in language learning and teaching. Moreover, evaluation and grading systems of AI technologies have shown their usefulness throughout the history of the CALL

studies. However, new technologies that arose recently such as AI enhanced chatbots promises even more impressive opportunities. Therefore, it is vital to test the reliability and consistency of these chatbots. Consequently, this section will analyse the grading of chatbots as the thesis's quantitative data. On a similar note, to the human grading; this analysis will include the all-essay types that were graded using TOEFL IBT Writing Evaluation Rubric. All chatbots are given the same prompt for all essays. They were provided with the rubric and were asked to evaluate a certain type of essay according to the rubric. Exact procedures in the human grading were applied in the AI grading as well.

3.4.1 Narrative Essays

Initial analysis of the grading in the narrative essays are quite surprising. Three different chatbots were used to grade essays. Shockingly, gradings were almost the same throughout. Apart from some 1-2 points difference, almost every grade for every criterion was the same. Consequently, most of the overall gradings were the same. Minority of them had differences that could be easily overlooked because the point differences was maximum two points. The gradings showed unexpectedly high consistency within each other. Mentioned differences will be better explained when evaluating the feedback of AI chatbots.

3.4.2 Expository Essays

Expository Essays indicate interesting results. Like other essay types, consistency of the chatbots is quite high for each criterion. Although some big differences can be seen between the chatbots, when compared to each other, consistency of the gradings are significantly high. The most consistent criterion for expository essays according to the analysis is the language use with only 0.15 rate of SD. Task response follow with 0.17, organisation with 0.37 and coherence with 0.47 SD ratio. What is interesting is that between criterion there is not even a 0.5 SD ratio, however overall gradings of the essay show a rather big difference contrary to the criteria. Overall grading shows 1.1 SD, which is still low but significantly higher than other criteria. In conclusion, other than the fact that the mean of overall grading is slightly high with more than 1 SD, all other criteria are showing significant consistency amongst each other indicate that the gradings of the essays are similar for chatbots in terms of expository essays.

3.4.3 Argumentative Essays

Analysis of the argumentative essays is slightly different than especially the narrative essays. In narrative essays the gradings were almost the same through the essays for each criterion. However, gradings of the argumentative essays show some different gradings for the essays. Before getting to analysis there is a slight issue that is worth mentioning. Gemini had graded all essays except one exactly the same with task response and organisation as 4 points, coherence and language use as 3 points each, and eventually overall grading as 14. However, for the fifth essay it graded task response as 4 and others as 3 which leads a total of 13. This of course when compared to other chatbots affect the end result of inter-rater consistency. Leaving Gemini aside; ChatGPT and BingAI showed great consistency for each criterion of each essay. Other than a few essays, gradings show a huge similarity with a difference of 0,5 between criterion and maximum 1,5 in overall grading. Greatest consistency in gradings found in task response with 0.23 SD. Least amount of consistency can be seen in coherence with 0.35 SD. Considering all, even the least consistent criterion is significantly consistent. Moreover, overall grading show only 0.8 SD which again shows consistent grading throughout the essays. The other criteria fall between 0.27 and 0.35 which is still considered as consistent enough. Even though the data from the Gemini is the same for all essays, apparently this does not lead to big problems since the gradings are favourably consistent especially even more so than human gradings.

3.4.4 Conclusion

In summary, narrative essays, expository essays, and argumentative essays are different essay types that are evaluated differently but with the same rubric. Three chatbots; ChatGPT, BingAI, and Gemini when asked to evaluate these essays using the TOEFL IBT Writing Evaluation rubric have all shown great promise in terms of successfully grading the essays. Analysis of the gradings of three chatbots provided invaluable data regarding especially the consistency of the chatbots also the accuracy of chatbots. The data clearly indicates significantly high consistency between each essay type and every criterion. Consequently, the results strongly suggest that grading done by chatbots are consistent with each other and has high accuracy. Of course, the trustworthiness and correctness of these gradings will be further explored in the comparison part.

3.5 Analysis of AI Feedback

The feedback of chatbots is slightly different than the grading for some reasons. First of all, although seems similar at first look a detailed examination reveals that chatbots have different ways of giving feedback. Especially for narrative essays BingAI and Gemini are somewhat consistent with the feedback they provide however, they show extremely repetitive results with little variety and personalized feedback. Generally, they praise the work of the student without getting into too much detail especially for some criteria such as task response yet when asked they are able to provide further personalized feedback which again can show some repetition. Here are two examples from the BingAI's feedback to two different narrative essays.

"1. Task Response: 4/5

The essay effectively addresses the task of narrating a personal experience. The writer provides a clear and detailed account of a specific day when things went wrong, including the events, their consequences, and the lessons learned. However, there could be a bit more depth in the reflection to fully meet the highest standards."

1. Task Response: 4/5

The essay effectively addresses the task of narrating a personal experience related to social anxiety. The writer provides a clear account of their struggles and how they overcame them. However, the essay could benefit from more detailed reflections on the lessons learned and personal growth."

As can be seen, BingAI falls into repetition whilst evaluating and lacks giving deep feedback to essays.

On the contrary, ChatGPT does a substantially better job at giving feedback. ChatGPT rarely feels repetitive and even when it is repetitive it is not out of place. Moreover, for every essay ChatGPT provides personalised feedback, extensive suggestions on the specific criteria, and ideas for further improvement to students for each criterion. Not only that but ChatGPT does a great job at exemplifying from within the essays to further improve the strength and accuracy of feedback. For instance, misuse of grammar, vocabulary items are highlighted in the feedback and possible replacements, or correct versions are suggested.

Here is an example of feedback from ChatGPT that focuses on the task response criterion.

“1. Task Response:

- *Score: 4/5*
- *The essay adequately addresses the topic of living with mental health problems and explores different mental disorders such as Bipolar Disorder, PTSD, and Eating Disorders. It provides relevant explanations for each disorder, including their symptoms, causes, and treatment methods. However, the discussion is somewhat surface-level and lacks deeper insight or critical analysis. More elaboration on societal impacts or personal perspectives could further enhance the task response.”*

In summary, although they are not terrible and provides somewhat decent feedback; BingAI and Gemini falls behind of ChatGPT rather quickly in terms of giving feedback. Consequently, the analysis of the feedback will mostly revolve around the feedback from ChatGPT since it is the most consistent and accurate, needless to say; BingAI and Gemini will not be overlooked but provide a supportive role.

3.5.1 Narrative Essays

The analysis of feedback given by AI has provided invaluable insight and important data. One of the most important and stand out discovery was that all three of chatbots focused on the same aspects of each criterion with rather similar results. However, the difference between the feedback was on detailing and giving examples. To begin with similarities, all chatbots regardless of the grading provided a positive tone and positive feedback for essays. In addition to positivity, all of them strongly focused on giving suggestions more than mistakes. The other important similarity is to themes that served as pivot points for each criterion.

For task response main themes that were constantly referred are requirements which comments on the essays’ ability to reflect the main purpose of the essay type, lesson learning that addresses the supposed lesson after a certain event, and personal growth that focuses on the students’ perspective on the immediate impact and the long-term impact of the event in their lives.

Following that, organisation had even more themes that were focused on. The most referred theme was structure that was repeatedly found in all essays, additionally paragraphing, logical flow of the essay, transitional aspects,

progression, and reflection were highlighted in the feedback. In addition to these themes essays' ability to setting a scene was reflected in the organisational aspects of the essays.

Next is coherence where the themes were not as ample as others. Coherence is evaluated and assessed under three main themes which is connection that refers to the essays way of connecting paragraphs or sentences with correct and appropriate use of linkers, descriptive language that addresses the vividness of essay, smooth transitions, and the consistency of essay that highlights the flow of the events.

Lastly for language use, the main themes were the grammar, vocabulary items, phrases, and sentence structures. All of the sub-themes revolved around these main themes and were interconnected. AI focused on the writers' competence to command the English language appropriately.

Additionally, overall feedback had the same structure for all essays but was different and individual. Basically, for overall feedback chatbots focused on rephrasing and summarising the entire feedback processes under one paragraph and added encouraging notes to students.

Besides similarities there were notable differences between the feedback given by chatbots. For narrative essays, although Gemini and ChatGPT had significantly more similarities, still BingAI and Gemini were behind ChatGPT in terms of few aspects. The biggest difference between ChatGPT and other chatbots were ChatGPT's individualistic and personalised feedback. ChatGPT provided ample examples from the essay and explained the suggestions in detail, whereas BingAI did not provide any examples from the essays. Gemini was inconsistent with the exemplifying. Here are two examples from ChatGPT and BingAI taken from the feedback of same essay's same criterion. The essay and the criterion were selected randomly, which are the language use part of the fifth essay.

"4. Language Use: 3/5

The language use is generally effective, with a good range of vocabulary and sentence structures. There are several grammatical errors and awkward phrasings that slightly impede understanding. The writer demonstrates a solid grasp of English, but there is room for improvement in accuracy and fluency."

This is the feedback from BingAI.

“4. Language Use:

Score: 2.5/5

There are several language issues that affect clarity, including incorrect word choices, grammar mistakes, and awkward sentence structures. For instance, phrases like "uniforms is considered" should be "uniforms are considered," and "provide equality" should be "provides equality."

Furthermore, punctuation is inconsistent, and sentences could be smoother. The essay would benefit from more precise vocabulary and greater attention to sentence variety. Additionally, the informal tone in parts (e.g., "kids hate them") detracts from the overall academic style.”

This is the feedback from ChatGPT. Obviously, although the grades are similar ChatGPT provided examples from the essay that enhances the quality of feedback on the contrary, BingAI had no examples and plain feedback without adding any details.

Additionally, another huge difference was the feedback for further improvement that was provided by ChatGPT and occasionally Gemini. Although the same prompt was used, BingAI did not provide any additional feedback that addresses the missing points and skills of writers based on the written narrative essays. Here is an example of “suggestions for further improvement” that was taken from the fifth narrative essay’s feedback of ChatGPT.

“Suggestions for Further Improvement:

1. Develop a Clear Thesis Statement:

- *Start the essay with a clear thesis that outlines the main points you will discuss. This gives the essay focus and provides a roadmap for the reader. For example: "This essay will explore both the benefits and drawbacks of school uniforms, particularly in relation to identity, equality, and student comfort."*

2. Strengthen Paragraph Structure and Transitions:

- *Organize the essay into clear paragraphs, each focusing on one specific point (e.g., one paragraph on equality, another on comfort, etc.). Use transitions between paragraphs to improve the flow of the essay and guide the reader through your argument.*

3. Improve Language Accuracy:

- *Focus on eliminating grammatical errors and awkward phrasing. For example, ensure subject-verb agreement, proper use of articles, and correct word forms. Consider varying sentence structures to make the writing more engaging and formal.*

4. Enhance Analysis:

- *Go beyond stating the pros and cons. Provide deeper analysis of how school uniforms affect students' mental well-being and learning environment. Discuss the long-term impacts or provide examples to strengthen your argument.*

5. Conclusion:

- *Add a conclusion that summarizes your points and restates the thesis. This will help tie the essay together and leave a lasting impression on the reader.*

Improving these areas will enhance the clarity, structure, and depth of the essay, helping you better convey your argument in a well-organized and grammatically accurate manner”.

ChatGPT provides further information and additional details for students to improve their writing skills. It is worth noting that these suggestions are individualised based on the essays.

In conclusion, analysis of the feedback for narrative essays that was done by chatbots reveals critical findings. According to the data, all chatbots provide somewhat decent feedback and successfully responds to the given prompt. However, ChatGPT proves to be superior chatbot out of three chatbots that were used in this project, at least for narrative essays. The reason for this is that ChatGPT gives better feedback by providing more details, on point examples, and extra suggestions. All the aspects considered, all chatbots are able to provide feedback to a degree however, ChatGPT carries it a few steps further than the other chatbots.

3.5.2 Expository Essays

Expository Essays as explained numerous times in previous paragraphs, offer different promises. “Explaining” a topic whilst preserving a neutral position might prove a difficult task for writers. Consequently, the way in which expository essays are assessed are slightly different for AI as it was for humans as well. Moreover,

this situation results in some different themes and sub-themes being found in the feedback.

Similarly to the narrative essays, expository essays had the same themes and sub-themes in all chatbots yet again the delivery is slightly different as well. However, for expository essays there are not huge differences in feedback. In narrative essays, especially BingAI fell short in giving feedback than other chatbots and ChatGPT proved to be superior. Although ChatGPT still better than other two chatbots, this time BingAI and Gemini are a lot better than they were in narrative essays.

To begin with, requirements was one of the main focal points of task response. Following requirements, clear statement was a significantly highlighted aspect of task response. Another aspect was the depth of essays which refers to the detailed exploration of the topic at hand. Last major feature of task response was the essays' ability to successfully provide an overview of the issue.

After that, organisation focuses on logical flow of ideas, structuring, paragraphing, and transitions. Additionally, the feedback emphasises the importance of how each paragraphs tackles an aspect of the issue.

Subsequently, similar themes are found in coherence as well. Naturally logical connection of paragraphs and flow are highlighted. Additionally, essays' inner consistency is challenged and commented on. For instance, connections between the first and the third paragraph are questioned on compatibility.

Finally, language use pivots around grammar and phrases as well as the range of lexical items. In addition to those which are similar to narrative essays, the tone and academic language is strongly pointed on for expository essays. Also, the conciseness and preciseness of the language is explored. Lastly, fluency and accuracy of the essays are confronted referring to the proficiency levels of learners.

It is worth noting that overall feedback of the expository essays served a similar role to expository essays as in narrative essays. General notes and errors are summarised with constructive suggestions. Moreover, students' proficiency levels are always considered and mentioned throughout the feedback at every possible turn.

Although the chatbots were closer in terms of accuracy this time. ChatGPT is still a few steps ahead of the other chatbots. The reasons are similar, it is more detailed, on point, and specific. Furthermore, it has significantly more concise and

precise commentaries about the essays. Subsequently, ChatGPT simply provides more examples from within the essays. Additional feedback is always given as like a bonus. However, this time all three of the chatbots provided somewhat detailed feedback. Here are the three additional feedback from ChatGPT, BingAI, and Gemini with exact order.

Suggestions for Further Improvement:

1. Add Specific Examples and Evidence:

- *To strengthen the essay, include concrete examples, such as key battles, major events (like the Bonn Agreement or NATO's involvement), or the effects on specific groups (such as women and children in Afghanistan). This will provide depth and support your arguments.*

2. Expand the Analysis:

- *You can improve your analysis by elaborating on the long-term consequences of the war, not only for Afghanistan but also for global politics. For example, how has the war influenced US foreign policy, or what are the regional implications in neighboring countries like Pakistan?*

3. Use More Sophisticated Vocabulary:

- *Try to expand your vocabulary and use more advanced sentence structures. Instead of simpler phrases like "international interest waned," you could say, "Global attention shifted away from Afghanistan as geopolitical priorities evolved." This would enhance the academic tone of your writing.*

4. Develop Ideas Further:

- *Some points in your essay, such as the political instability and societal challenges in Afghanistan, could be explored in more detail. Explain how the war has impacted governance, education, or the economy in Afghanistan, providing a more comprehensive view of the conflict's consequences.*

5. Strengthen the Conclusion:

- *The conclusion could offer a stronger, more reflective final thought on the significance of the war. What are the lessons learned for international diplomacy? How might this shape future global*

interventions? Offering a broader perspective will leave a lasting impact on the reader.

By incorporating these improvements, future essays can demonstrate a higher level of critical thinking, analytical depth, and language proficiency. Keep practicing, and you'll see steady progress!

Detailed Feedback for Improvement:

- *Introduction: Start with a clear thesis statement that outlines the main points of the essay.*
- *Body Paragraphs: Ensure each paragraph has a clear topic sentence and directly supports the thesis statement. Use more varied sentence structures and advanced vocabulary.*
- *Conclusion: Summarize the key points discussed in the essay and provide a strong closing statement that reinforces the importance of understanding the Afghanistan War's legacy.*
- *Language Use: Review for minor grammatical errors and awkward phrasings. Aim for concise sentences and avoid repetition.*

Keep up the good work! With these improvements, your writing will become even more effective and engaging.

Suggestions for Improvement:

- *Expand on the consequences: Explore the broader implications of the war, including its impact on regional stability, global counterterrorism efforts, and the humanitarian crisis.*
- *Incorporate different perspectives: Include perspectives from Afghans, women, and minorities to provide a more nuanced understanding of the war's impact.*
- *Analyze the factors contributing to the war's longevity: Discuss the reasons why the conflict persisted for so long, including the complex dynamics of the region, the challenges of nation-building, and the role of external actors.*
- *Consider the future of Afghanistan: Explore the challenges facing Afghanistan in the post-war era, such as political instability, economic development, and reconciliation efforts.*

By incorporating these suggestions, the essay could be strengthened to provide a more comprehensive and insightful analysis of the Afghanistan War.

In conclusion, on the contrary to narrative essays, the feedback from expository essays provided data where the chatbots were slightly more similar in terms of performance. Needless to say, analysis suggests that chatbots are effective tools whilst evaluating and giving feedback to expository essays. Still worth mentioning that ChatGPT is better but other chatbots are able to quickly respond to the improvements.

3.5.3 Argumentative Essays

Argumentative Essays are similar to expository essays in terms of structure and language. Consequently, they show similarities in feedback as well. The main difference between argumentative essays and expository essays is that in argumentative essays writers have a position, idea, or an argument that they are for or against. This main difference reflects to the feedback. Overall, themes and sub-themes that were found in argumentative essays show parallelism to those found in expository essays.

In terms of performance of the chatbots, it can be seen that the performance is relatively closer to expository essays where the chatbots had similar results with similar accuracy in feedback. However, once again ChatGPT performs better than other chatbots due to the similar reasons that were revealed in expository essays and narrative essays.

Regardless of the performance, all chatbots had common themes in each criterion. For task response, the most frequently mentioned criterion was requirements and having successful argumentation. Supporting ideas and comprehensive explorations of the issues were highlighted in the feedback. Also, the positions of the writers and how well they hold those positions are criticized. The last important aspect was the writer's ability to analyze the problems, counterarguments, and different perspectives within the essay providing appropriate responses and examples.

As for organization, logical structuring is found to be an important pivot for argumentative essays. Transition and connection of different paragraphs were stressed, and smoothness of the essays were appraised. Structure and paragraphs are emphasized to ensure clearer organization.

Followingly, coherence had rather similar themes to the other two essays. Flow and smoothness were two main focal points whilst assessing coherence.

However, this time clarity and length of the sentences, and repetition of ideas or arguments were highlighted and criticized in most of the essays.

Lastly, language use mainly focused on the issues regarding the vocabulary that was used, grammatical mistakes and errors, and phrasings that were used. Additionally, essays' tone and the use of academic language were strongly emphasized. On a similar note, the overall feedback provided a general summary of the given feedback mentioning all the most important mistakes, faults, or criticized points in the essay.

To sum up, although especially for BingAI and Gemini there could be clear improvements to enhance the results of the feedback, the performance of chatbots and their accuracy proved to be useful once again. Additionally, chatbots' ability to realise argumentation and criticizing shown to be effective. On the whole, the data indicated that chatbots are effective tools to address appropriate feedback to students whilst evaluating argumentative essays.

3.5.4 Conclusion

In conclusion, feedback is an important tool to address the students' needs, and it has been proven that quality feedback plays a critical role in improving the desired skills of the students. Written feedback that concentrates on the students' writing skills is important for better learning and improvement. This study tested the efficiency and effectiveness of three different chatbots in terms of providing quality feedback to students on three different essay types. The data indicated that chatbots are superefficient assets in terms of providing feedback. However, there are notable differences in the performances of three chatbots. The results revealed that although all three of the chatbots provide decent feedback, ChatGPT is considerably more efficient and better as a result of providing more individualistic and detailed feedback enriching with on point examples. The data also revealed that chatbots can effectively utilize a provided rubric with consistency regardless of the essay type. In this study, TOEFL IBT Writing Evaluation rubric was sufficiently used by chatbots. This resulted in consistent results even for different essay types. Chatbots successfully used the rubric, created general themes for all essay types moreover, they created specific themes and sub-themes for each essay type. For instance, the vividness of the essay was considered to be an important aspect for narrative essays whereas strong argumentation and academic language were highlighted in the argumentative essays. Overall, although some chatbots may

perform slightly under than others, the data strongly suggest that chatbots, and especially ChatGPT, are effective assets in terms of providing quality, effective, consistent, and user-friendly feedback.

3.6 The Comparison Between Human and AI

After collecting and analyzing the two datasets in detail, it is critical to find possible similarities and differences and to compare human datasets and AI datasets in order to figure out how effective can AI be in terms of grading and giving feedback. It goes without saying that different teacher/evaluator groups, different essay types, or different rubrics might slightly affect the results. However, for this particular group of teachers, essays, and essay types; in this section a detailed comparison of the grading and feedback of humans and AI chatbots will be explored. First of all, the gradings will be compared by comparing the mean numbers, standard deviations, and end results. This will be done separately for each essay type to get a better understanding of the probable effects of different essay types on grading. Second of all, the feedback from teachers and AI chatbots will be investigated. The investigation will delve into the feedback by examining the feedback from length to complexity. Certain aspects like the tone and language of the feedback will be explored. The results of the analysis will provide a clear insight into the practicality of AI chatbots in terms of grading and giving feedback to three different types of essays.

3.6.1 The Comparison of Gradings

Grading is a significantly important part of evaluation and assessment. It provides learners and teachers with a glimpse of the learners' performance and their closeness to the ideal performance. Grading processes are affected by several aspects. However, to minimize the possible effects of these aspects some limitations are made. Rubrics serve as a great guide to create an overall path for teachers/evaluators to follow reducing the mentioned effects. Following a rubric makes it easier to ensure a degree of consistency through different essays and essay types. Although that is the case, it is natural to encounter differences between different evaluators, especially if one of the evaluators happens to be machine. Consequently, this part will compare the gradings by human evaluators and chatbots. This comparison will first focus on the overall gradings in detail to define the overall gradings reflection on both groups.

Before all, a normality test was conducted to determine the spread of the overall gradings. For each essay type a normality test was performed. The normality test focused on the average overall gradings of human evaluators and examined chatbots' grading individually.

Table 3.1 Normality Test of Narrative Essays

Test

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
HGAvg	.141	10	.200*	.934	10	.489
ChatGPT	.189	10	.200*	.908	10	.269
BingAI	.375	10	<.001	.804	10	.016
Gemini	.412	10	<.001	.647	10	<.001

*. This is a lower bound of the true significance.

^a. Lilliefors Significance Correction

Followingly, the results of the normality tests were interesting. Firstly, in narrative essays both teachers and chatbots showed a relatively normal spread. Out of all four values, ChatGPT and average of teachers had the lowest skewness value; BingAI and Gemini had slightly higher skewness indicating slight deviation from normal. However, their kurtosis levels indicated a relatively normal spread in the gradings. Considering Shapiro-Wilk numbers, W value showed W=0.934 for teachers, W=0.908 for ChatGPT, W=0.804 for BingAI, and W=0.647 for Gemini. Additionally, teachers had a p value of p=0.48, ChatGPT p=0.269, BingAI p=0.016, and Gemini had a p of p=<0.001. Consequently, The Shapiro-Wilk test of normality indicated that the teachers and ChatGPT were approximately normal resulting in the failing of the null hypothesis of normality. On the other hand, the test indicated a considerable deviation from normality for BingAI and Gemini suggesting that the data is not normally distributed. The table below shows the results of both The Shapiro-Wilk test of normality and The Kolmogorov-Smirnov test for narrative essays.

Following narrative essays, expository essays underwent the same process. Before beginning to compare the results, a normality test was conducted to find out

the spread of the grades. The results of the normality test for expository essay showed differences than narrative essays for all parties. First of all, the skewness and kurtosis values were significantly different. Skewness values were higher than ± 1 indicating high skewness; however, kurtosis levels were closer to normal ranges. Followingly, the values of Shapiro-Wilk test indicated that the most normally distributed data belonged to ChatGPT with $W=0.888$ and $p=0.160$, the other normal distribution was the human averages with $W=0.862$ and $p=0.081$. Gemini was almost normal with $W=0.813$ and $p=0.021$ and BingAI was totally non-normal with $W=0.594$ and $p<0.001$. Although Gemini was quite close to being normally distributed with its W value, p must be larger than 0.05 for normal distribution. Below is a table of the results of Kolmogorov-Smirnov and Shapiro-Wilk tests of normality for expository essays.

Table 3.2 Normality Test of Expository Essays

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
HGAvg	.222	10	.178	.862	10	.081
ChatGPT	.267	10	.042	.888	10	.160
BingAI	.433	10	<.001	.594	10	<.001
Gemini	.216	10	.200*	.813	10	.021

*. This is a lower bound of the true significance.

^a. Lilliefors Significance Correction

Last of all, the results of the normality test of the argumentative essays show considerable differences than the other two essay types. To begin with skewness and kurtosis values, especially BingAI and teachers have normal skewness and kurtosis; additionally, ChatGPT shows moderate skewness but normal kurtosis, however Gemini shows significantly high skewness and heavy tails. Considering Shapiro-Wilk results ChatGPT again proves to be the most normally distributed data with W and p values being $W=0.951$ and $p=0.678$, different than other results BingAI follows ChatGPT with $W=0.913$ and $p=0.3$, human average follow BingAI closely with $W=0.901$ and $p=0.223$. The results indicate that for argumentative essays, ChatGPT, BingAI, and Teachers have normal distribution. However,

Gemini with $W=0.366$ and $p<0.001$ have non-normal values. The table below shows the results of The Kolmogorov-Smirnov and The Shapiro Wilk tests of normality for argumentative essays.

Table 3.3 Normality Test of Argumentative Essays

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
HGAvg	.155	10	.200*	.901	10	.223
ChatGPT	.181	10	.200*	.951	10	.678
BingAI	.152	10	.200*	.913	10	.300
Gemini	.524	10	<.001	.366	10	<.001

*. This is a lower bound of the true significance.

^a. Lilliefors Significance Correction

In conclusion, the normality test was conducted before comparing the results to identify the normalness of the data. The data is critical to define on accounts for explaining normality to further investigate the comparison by the results suggesting the correct choice of further tests. According to the findings of the normality tests ChatGPT showed the biggest consistency and had normal distribution for all essay types. Similarly, teachers had normal distribution but lacked flying numbers in expository essays. BingAI had only normal distribution in argumentative essays and Gemini did not have any normal distribution for any essay type.

Following the results of the normality tests, One-Way Repeated Measures Test (ANOVA) or the Freidman will be used for comparison. ANOVA will be used between the groups that reject the null-hypothesis and prove to be normally distributed. Additionally, the Freidman test will be utilized for non-normal groups.

According to the results that the normality test showed, the first comparison was made using ANOVA between the average of teachers and ChatGPT for narrative essays followed by expository and argumentative essays. ChatGPT and Teachers had a normal distribution for all essay types. For narrative essays the results of the ANOVA test indicate a relatively high F value, but p value does not indicate any significant difference. In short, the results show that while there is

variability among groups, considering the $p=0.074$ therefore $p>0.05$, the variability is not enough to reach statistical significance under the conditions of this particular study. Subsequently, almost the same results are found in the analysis of expository essays with a relatively high F value, and $p=0.072$ meaning that similar to narrative essays, the analysis of the expository essays did not find meaningful difference amongst the two groups. Lastly, the analysis for argumentative essays found the p value of $p=0.063$ with a similar F value to narrative and expository essays. In conclusion, the findings of the ANOVA test that was conducted to determine possible significant differences between the average grade of teachers and the grades of ChatGPT failed to find significant differences between the data with all of them having larger p value than 0.05, indicating that they are not significantly different in terms of grading. (Below are the tables for ChatGPT vs. teachers)

Table 3.4 Narrative Essays ANOVA Teachers vs. ChatGPT

Measure: MEASURE_1

Source	ChatGPthuman	Type III Sum of Squares	df	Mean Square	F	Sig.
ChatGPthuman	Linear	8.889	1	8.889	4.074	.074
Error(ChatGPthuman)	Linear	19.639	9	2.182		

Table 3.5 Expository Essays ANOVA Teachers vs. ChatGPT

Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	chatgpthuman	Type III Sum of Squares	df	Mean Square	F	Sig.
--------	--------------	-------------------------	----	-------------	---	------

chatgpthumn	Linear	12.482	1	12.482	4.153	.072
Error(chatgpthumn)	Linear	27.048	9	3.005		

Table 3.6 Argumentative Essays ANOVA Teachers vs. ChatGPT

Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	chatgpthuman	Type III Sum of Squares	df	Mean Square	F	Sig.
chatgpthuman	Linear	9.385	1	9.385	4.504	.063
Error(chatgpthuman)	Linear	18.751	9	2.083		

The second comparison of grading was conducted between the average of teachers and BingAI. The results of the normality tests showed that only normal distribution in BingAI was argumentative essays. Therefore, to analyze the possible differences, the Freidman test was made for narrative and expository essays, and ANOVA for argumentative essays. The findings of the Freidman test that was conducted for narrative essays did not find any significant difference between the gradings with $p=0.527$. Interestingly, with an exact p value, the analysis of the expository essays did not indicate significant differences as well. However, according to the One-Way Repeated Measures test that was conducted for the analysis of argumentative essays detected significant differences between teachers and BingAI in the grading of argumentative essays with p value being $p<0.05$. In summary, the tests that were conducted to compare BingAI and teachers showed interesting results. While the results failed to reject the null hypothesis for narrative and expository essays, it was found that there were significant differences between

the groups in the grading of argumentative essays. (Below are the tables for BingAI vs. teachers)

Table 3.7 Narrative Essays Friedman Teachers vs. BingAI

<i>Test Statistics^a</i>	
N	10
Chi-Square	.400
df	1
Asymp. Sig.	.527
a. Friedman Test	

Table 3.8 Expository Essays Friedman Teachers vs. BingAI

<i>Test Statistics^a</i>	
N	10
Chi-Square	.400
df	1
Asymp. Sig.	.527
a. Friedman Test	

Table 3.9 Argumentative Essays Friedman Teachers vs. BingAI

Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
bingaihuman	Linear	35.644	1	35.644	31.951	.000
Error(bingaihuman)	Linear	10.040	9	1.116		

Lastly, for the comparison of Gemini and teachers, only the Freidman test was conducted for all essay types since Gemini did not have any normal distribution found. The first Freidman test which was for narrative essays found $p=0.527$, the same as the results for BingAI. Similar to BingAI, the value does not indicate any significant difference. However, the very next Freidman test indicated a significant difference between Gemini and teachers for expository essays due to p value showing $p=0.011$. Lastly, the analysis of the argumentative essays showed that there were no significant differences between the groups since p being $p=0.058$. Summary of the data shows that Gemini had no significant differences in terms of grading for narrative and argumentative essays whereas the findings indicated significant differences in the grading expository essays. (Below are the tables for Gemini vs. teachers)

Table 3.10 Narrative Essays Freidman Teachers vs. Gemini

<i>Test Statistics^a</i>	
N	10
Chi-Square	.400
df	1
Asymp. Sig.	.527
a. Friedman Test	

Table 3.11 Expository Essays Freidman Teachers vs. Gemini

<i>Test Statistics^a</i>	
N	10

Chi-Square	6.400
df	1
Asymp. Sig.	.011
a. Friedman Test	

Table 3.12 Argumentative Essays Friedman Teachers vs. Gemini

<i>Test Statistics^a</i>	
N	10
Chi-Square	3.600
df	1
Asymp. Sig.	.058
a. Friedman Test	

All things considered, the results of the analysis of normality tests and comparison tests showed that amongst the AI chatbots, ChatGPT was closer to humans and comparison tests found no significant difference in all essay types. However, while BingAI had significant differences in argumentative essays, Gemini had significant differences in expository essays. In terms of normality, ChatGPT even surpassed teachers at some point and had more normally distributed data. The end results indicate that while BingAI and Gemini had differences from teachers at certain essay types they are still valuable and had no significant differences for other essay types, furthermore ChatGPT had no significant differences than teachers for all essay types showing that it is considerably close in terms of grading to teachers.

Following the analysis of overall gradings, more specific and in-depth analysis was made to take a closer look at the numbers in other criteria. First of all, narrative essays were the first essay type that was evaluated by both teachers and chatbots. A comparison within the individual groups was made beforehand and the analysis will consider those comparisons as an average of both groups. The analysis of the gradings indicate interesting results. For task response, the mean number of

teachers' grading is 4.08 with 0.13 SD, similarly AI grading show a flat 4 with 0.1 SD. Followingly, grading of organization of teachers present a mean grade of 3.6 with 0.17 SD, these numbers are 3.5 and 0.3 for chatbots. After that there is coherence with a mean of 3.8 with 0.24 SD for teachers, and 3.4 with 0.26 SD for chatbots. On a similar note, language use has a mean value of 3.4 and 0.32 SD for teachers and 2.8 with 0.32 for chatbots. According to the results, there are not huge differences on the minimum and maximum values in gradings. Similarly, the average gradings are also quite close, to further explain, for task response the difference stands to be 0.08 points, as for organization this number is 0.1, followed by the number for coherence that is 0.4, and language use shows a grade of almost 0.6 difference in grading. (See Appendix 3)

Secondly, the results of expository essays show slight differences than narrative essays. The analysis of the teachers and chatbots are as followingly. Initially, the task response presents that teachers have an average grading of 4.1 with 0.27 SD and chatbots have 3.98 and 0.17 SD. Afterwards the organization gives similar numbers with 3.84 and 0.4 SD for teachers in addition to 3.78 and 0.37 SD for chatbots. Coherence on the other hand has a faintly more difference, for teachers the numbers are 4.18, 0.27 SD and for chatbots 3.66, 0.47 SD. On a similar note, language use has 4.18 and 0.17 SD for teachers and 3.5 with a 0.15 SD for chatbots. Consequently, the differences for each criterion can be seen as follows; task response gives a total difference of 0.12, organization gives a slightly less number that is 0.06, as for coherence the number vaguely increases which is almost 0.5, followingly language use has a similar number to coherence with a difference of nearly 0.7. (See Appendix 3)

The last analysis will be gradings of the argumentative essays. To begin with, for teachers the mean grading of the task response stands at 4.4 with 0.5 SD, these numbers are 3.8 with 0.2 SD for chatbots. In the organization section teachers have an average of 3.63-0.8 SD and chatbots have 3.8-0.3 SD. Additionally, coherence presents an average of 4 with 0.4 SD for teachers and 3.55 with 0.35 for chatbots. And for language use teachers present the average number of 4 with 0.2 SD and chatbots present the average number of 3 with 0.2 SD. In conclusion, the differences for each criterion are 0.6 for task response, 0.17 for organization, 0.45 for coherence, and 1 for language use. (See Appendix 3)

To conclude, the data indicates that the least number of differences are found in the gradings of narrative essays followed by expository essays and relatively the biggest differences in argumentative essays. Considering the standard deviations, most of the differences can easily be overlooked to a degree to say that there are not even differences. Minimum and maximum values are all found to be similar in each essay type for each criterion. According to the results, especially in argumentative essays, teachers had more standard deviation than chatbots that vaguely affect reliability to a degree. However, the fact that both teachers and chatbots showing similar results indicates that the reliability and trustworthiness of both groups are significantly high. All things considered, the analysis of the data and the findings indicated that the chatbots and teachers performed considerably similar to each other and the differences in gradings are incredibly low and noticeably consistent throughout the essays and essay types. The data also indicated that chatbots are slightly more consistent in grading with less standard deviations. It is worth mentioning that chatbots when taken individually would alter the results depending on the choice of the chatbot. To sum up, the analysis of the data of this particular study indicates that the chatbots can be used as invaluable assets whilst grading essays since they give significantly similar results to teachers' gradings.

3.6.2 The Comparison of Feedback

The comparison of the feedback of the essays will consider the examining that was done before for both teachers and chatbots. The conciseness, preciseness, quality, and consistency of the feedback will be critical to compare the two groups. Moreover, the length of feedback, the suggestions and examples given in feedback, and the tone of the feedback will aid in the process of comparing teachers and chatbots and their performances of providing quality feedback.

Since the data is similar for each essay type; the comparison of given feedback will not be done separately. The data showed that teachers have variety of themes whilst evaluating each criterion of the essays. There were of course common themes, however teachers relied on their own criteria as well in the process of evaluation. On the other hand, all chatbots had the same themes and similar sub-themes for each criterion and they limited themselves to those specific themes and sub-themes. For instance, one of the teachers repeatedly focused on the titles of the essays, however no other teacher considered title as a factor in feedback. Here are

some examples of the teachers' feedback of the task response criterion of argumentative essays.

"The title and the content of the article are connected and convey the intended message effectively. 5/5"

"The essay sounds quite professional. It is fully responding the requirements of the task 5/5"

"5 - The essay is well-written with excellent coherence and cohesion, nicely organized into chunks of ideas and clearly puts forward an argument with a conclusion."

Another example is that chatbots often referred to the vividness of the essays in narrative essays, whereas teachers rarely considered vividness whilst giving feedback. This condition has advantages and disadvantages, one major down point is that it may affect the consistency of the feedback and inter-rater reliability since there are numerous themes that some teachers might find unrelated to the topic, on the other hand it might create new perspectives and provide some insight into other teachers' opinions. Moreover, teachers generally avoided making suggestions and sometimes some teachers had negative commentaries with negative tones, on the other hand even when the grading was low chatbots evaded negative tones and always tried to encourage students with ample suggestions. In addition to suggestions, especially ChatGPT provided numerous examples when needed from the essays; although teachers included examples in the feedback, there were lack of examples in the feedback of teachers. Below are samples from the organisation part of the same essay that was given feedback by two teachers and ChatGPT.

"The paragraphs do not logically follow the previous one, they are not connected."

"2- The essay is divided into two paragraphs, however, the body paragraphs are intertwined with random development. The development part does not have topic sentences and are not organized into single ideas. It is written in a stream of consciousness way and this organization is not suitable for an essay."

"The essay lacks a clear organizational structure. While it has an introduction, body, and conclusion, the points within each section are not always well-connected. The introduction sets up the topic, but the essay jumps between different forms of fanaticism—sports, religion,

etc.—without clear transitions. The essay would benefit from dividing the argument into distinct paragraphs for each form of fanaticism discussed (sports, religion, etc.) and using more signposting to guide the reader through the argument. The conclusion could also be clearer in summarizing the key points and reinforcing the overall message.”

The first to belong to teachers and the third one is from ChatGPT. Additionally, it is worth noting that, BingAI and Gemini rarely exemplified their feedback, so ChatGPT is a few steps ahead from teachers and other chatbots in terms of giving on point examples. Accuracy of the teachers and chatbots were relatively similar and especially ChatGPT was as concise and precise as teachers in providing quality feedback.

Followingly, it goes without saying that there are some advantages and disadvantages of both sides. First of all, teachers are humans and being human comes with certain drawbacks when compared to machinery. Teachers can be exhausted, tired, bored, sad, upset, or in many other emotional states. No matter how professional a teacher is, emotional state of the teacher might have positive or negative effect on giving feedback. Additionally, teachers generally are personally involved with the students, they know and recognise them, and this might lead to having biased opinions in the feedback. Moreover, evaluating, grading, and giving feedback to essays are long and tiring processes. It takes lots of times and might impede teachers from performing their actual duty more effectively, that is, teaching. Teachers are of course humans, and they are able to get a feeling of students' actual intent but there are numerous setbacks that might negatively affect teachers' well-being. On the other hand, chatbots are robots, machines. They do not get tired or exhausted nor they feel any emotion towards anything. Consequently, chatbots can give feedback to tens of essays in almost an instant objectively, and the data shows that they do it quite effectively.

In conclusion, the data and the comparison indicate that chatbots are at least as effective as teachers and sometimes they can do better jobs at giving feedback. Some setbacks that teachers face might affect the state of teachers and the feedback. However, only setback chatbots have is being less improved versions of themselves than tomorrow. The analysis shows that chatbots are great assets that teachers can utilise whilst giving feedback and the state of the technology suggests that they will only get better. However, it is important for teachers to use precise and accurate

prompts and always check the results for any possible errors and mistakes in feedback.

3.7 The Comparison of Chatbots

The results of the comparison between teachers and chatbots required further inspection for the potential differences regarding the chatbots. As previously discussed, ChatGPT and teachers had no significant differences in grading, whereas teachers and BingAI also Gemini showed significant differences for some essay types. The previous results indicated potential differences within chatbots, therefore this section will inspect and compare grading and feedback of chatbots within themselves.

Similar to the teacher-chatbot comparison, the results of normality tests decided the comparison test for chatbots. To begin with, ChatGPT and BingAI were compared using the Freidman Test for narrative and expository essays, and One-Way Repeated Measures ANOVA Test for argumentative essays. Firstly, the analysis of narrative essays showed the p value as $p=0.480$ which is considerably surpassing the accepted 0.05, indicating no significant differences between ChatGPT and BingAI. On the contrary, the p value was found to be $p=0.020$ for the comparison of expository essays, since $p<0.05$, the analysis found significant differences between ChatGPT and BingAI. Similarly, the p number for argumentative essays was below 0.05 with a value of $p=0.018$, also indicating significant differences between ChatGPT and BingAI in terms of grading. (Below are the tables for ChatGPT vs. BingAI)

Table 3.13 Narrative Essays Freidman ChatGPT vs. BingAI

<i>Test Statistics^a</i>	
N	10
Chi-Square	.500
df	1
Asymp. Sig.	.480
a. Friedman Test	

Table 3.14 Expository Essays Friedman ChatGPT vs. BingAI

<i>Test Statistics^a</i>	
N	10
Chi-Square	5.444
df	1
Asymp. Sig.	.020

a. Friedman Test

Second analysis was conducted between ChatGPT and Gemini. Since Gemini had no normal distribution for any essay type, only the Friedman test was conducted. According to the results, $p=0.739$ showed no significant differences between ChatGPT and Gemini in the grading of narrative essays. On a similar note, with a value of $p=0.527$ expository essays also showed no significant differences. In order to label the data as significantly different, p value needs to be $p<0.05$, consequently, differently from the other two, argumentative essays showed significant differences between ChatGPT and Gemini in terms of grading with a p value of $p=0.034$. (Below are the tables for ChatGPT vs. Gemini)

Table 3.15 Argumentative Essays ANOVA ChatGPT vs. BingAI

Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	Chat gptBing	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
ChatgptBing	Linear	8.450	1	8.450	8.40	.018	.483
Error(Chat gptBing)	Linear	9.050	9	1.006			

Table 3.16 Narrative Essays Friedman ChatGPT vs. Gemini

<i>Test Statistics^a</i>	
N	10
Chi-Square	.111
df	1
Asymp. Sig.	.739

a. Friedman Test

Table 3.17 Expository Essays Friedman ChatGPT vs. Gemini

<i>Test Statistics^a</i>	
N	10
Chi-Square	.400
df	1
Asymp. Sig.	.527

a. Friedman Test

Table 3.18 Argumentative Essays Friedman ChatGPT vs. Gemini

<i>Test Statistics^a</i>	
N	10
Chi-Square	4.500
df	1
Asymp. Sig.	.034

a. Friedman Test

Last analysis was to examine the BingAI and Gemini for potential differences. For the comparison, the Freidman test was conducted regarding the results of the normality tests. The results of the Freidman test indicated that there were no significant differences between BingAI and Gemini for all essay types. The results showed the p values as $p=0.206$ for narrative essays, $p=0.058$ for expository essays almost getting below the threshold, and $p=0.408$ for argumentative essays, showing BingAI and Gemini had no considerable differences for grading. (Below are the tables for BingAI vs. Gemini)

Table 3.19 Narrative Essays Freidman BingAI vs. Gemini

<i>Test Statistics^a</i>	
N	10
Chi-Square	1.600
df	1
Asymp. Sig.	.206
a. Friedman Test	

Table 3.20 Expository Essays Freidman BingAI vs. Gemini

<i>Test Statistics^a</i>	
N	10
Chi-Square	3.600
df	1
Asymp. Sig.	.058
a. Friedman Test	

Table 3.21 Argumentative Essays Freidman BingAI vs. Gemini

<i>Test Statistics^a</i>	
N	10

Chi-Square	.500
df	1
Asymp. Sig.	.480

a. Friedman Test

In conclusion, the results of the comparison of gradings indicated interesting results between chatbots. The findings indicated that for ChatGPT and BingAI, there were significant differences in expository essays and argumentative essays. As for ChatGPT and Gemini there were only significant differences in argumentative essays and no differences has been found in narrative and expository essays. Lastly, according to the results no significant differences has been found between BingAI and Gemini for narrative essays, expository essays, and argumentative essays.

As for feedback, the analysis followed similar procedures when comparing teacher feedback to chatbot feedback. However, only chatbots were considered this time. According to the analysis of the feedback from ChatGPT, BingAI, and Gemini, no significant differences were found between the themes that chatbots used. For all essay types chatbots used similar themes, such as, “vivid descriptions” for narrative essays, or “strong argumentation” for argumentative essays. In short, on the surface there was no considerable differences. However, more detailed examination of feedback revealed ChatGPT to be more successful due to some reasons that were mentioned above. ChatGPT provided more detailed and precise feedback, supporting its comments with multiple points and exemplifying using passages from the essays. Moreover, ChatGPT provided more detailed suggestions for students avoiding repetitive comments within and between the essays. BingAI and Gemini showed clear repetitive comments between the essays slightly affecting the reliability of chatbots. The analysis clearly showed that although chatbots perform similar in terms of feedback using almost identical themes, ChatGPT outperformed both BingAI and Gemini in terms of providing high quality feedback.

In conclusion, chatbots showed promising results providing high quality feedback and especially ChatGPT performing almost on-par with the teachers, and despite of their lacks, BingAI and Gemini are able to provide effective quality

feedback to different types of essays, slightly underperforming when compared to ChatGPT. The results clearly indicate that chatbots can be considered as effective tools in terms of feedback.

3.8 Conclusion

All things considered, the analysis of the data and the comparison show that chatbots, especially ChatGPT, are efficient and successful tools that teachers and evaluators can benefit from in terms of evaluation and assessment of an essay whilst grading and giving feedback. Teachers with highly trained eyes and necessary experiences are still vaguely above chatbots in terms of performance but even now this difference can easily be overlooked. Moreover, it will soon change directions in the opposite way quickly and chatbots will overperform teachers. However, as of the time of this study, teachers should always take the data with a grain of salt and double, perhaps triple check the results of chatbots. This will not only help teachers to get a load off their shoulders but also help learners because they can control the process and do it at home by themselves. Additionally, the findings showed considerable differences between chatbots as well both in grading and in feedback. Considering ChatGPT performed better than BingAI, and Gemini, AI literacy gains an important consideration not only for checking the information from chatbots but also being able to choose the right chatbot for the right job. However, as numerous mentioned, all chatbots are in a position that can mostly perform tasks efficiently. In summary, the analysis of the data strongly indicates that the chatbots are in a place that can efficiently evaluate and grade the essays with great consistency, accuracy, and efficiency. Moreover, they are able to provide quality feedback with great suggestions and examples for learners to improve their writing skills. They are at least on-par with teachers in terms of performance. Furthermore, they are significantly faster than teachers and easy to use. However, as aforementioned teachers, evaluators, or users should be considerate and careful whilst using chatbots. Teachers can carefully place their trust in the chatbots to aid them whilst grading and giving feedback to learners.

4. DISCUSSION

Although Artificial Intelligence has a history in grading and even in feedback, AI enhanced chatbots are relatively new technologies. OpenAI has published their multi-modal enhanced chatbot ChatGPT 3.5 in 2022. Since then, the uprising of AI enhanced multi modal technologies has become a mainstream topic for large companies. Following the success, Microsoft has released BingAI in February 2023, and Google has released Gemini in December 2023. These fresh technologies have grown both in popularity and capability drastically. In no more than two years they became one of the greatest technological advancements in the history. However, since they are this new, the importance and potential of chatbots were discovered even later. Many fields tend to integrate newer AI technologies in their studies for almost a year. English language teaching as a field is not different, the studies revolving around ChatGPT, and other AI enhanced multi-modal chatbots do not date back to earlier than 2023. Therefore, there is a lack of studies that discusses the abilities, opportunities, and usefulness of chatbots in ELT and more specifically in grading and providing quality feedback to different types of essays. In this part, the results of this particular study will be interpreted and compared to the recent studies that have similar goals to better investigate the data and to discuss the findings by exploring other studies. Consequently, the discussion part will provide us with the meaning and importance of this study whilst addressing the literature.

Reminiscently, the study has several key findings for grading and feedback. These findings are mostly related to the comparison between AI and teachers, mentioning the comparison between chatbots as well. Before getting to each criterion, the first analysis was to figure out the normality of the average of teachers and chatbots alike in terms of their performance on the overall grading. The results of this process showed interesting results. According to the findings, ChatGPT and teachers had normal distribution of data throughout each essay type, there were even moments where ChatGPT had more “normal” results than teachers. BingAI and Gemini failed to catch up to the other two groups in terms of normality with BingAI having only one normal and Gemini having zero normal distribution. After the normality, a comparison was made using either One-Way Repeated Measures Test (ANOVA) or the Freidman test depending on the normality results. Following

comparison tests the results indicated that ChatGPT and teachers had no significant differences in all essay types, whereas BingAI and teachers had significant differences in argumentative essays and Gemini and teachers had significant differences in expository essays. Naturally, within themselves chatbots had significant differences as well. Specifically, ChatGPT and BingAI had significant differences in expository essays and argumentative essays, ChatGPT and Gemini had considerable differences in argumentative essays, however the results found no significant differences in terms of grading between BingAI and Gemini. The comparison of the overall gradings showed that out of all chatbots that were used in this study, ChatGPT was the closest to teachers with no difference being found in any essay type. Additionally, although BingAI and Gemini failed to reach to the potential success of ChatGPT, they still had no significant differences in two different types of essays to teachers which can be considered as decent results.

After the analysis of overall gradings, more specific observation was made within each criterion. First of all, the results indicated that in terms of grading for all criteria, chatbots are consistent and accurate within themselves, additionally, chatbots perform on par with teachers and the gradings are significantly similar to each other. Differences can be easily overlooked because a difference between a chatbot and a teacher are nearly the same as a difference between two teachers grading the same essay. Secondly, as for feedback neither group had a better performance than the other. Teachers had some advantages, chatbots had other advantages. For instance, chatbots were fast as they gave elucidated quality feedback literally in an instant. They were more focused on the critical aspects of the rubric, they could give precise examples and great suggestions, and they were brilliant in detecting the grammatical mistakes. On the other hand, teachers had more nuanced and broader perspective and performed slightly better at noticing abstract concepts. Overall, according to the results it can be suggested that chatbots are now at a place that can be considered as almost equals to teachers in terms of grading and providing quality feedback to students.

The findings of the study indicated that the chatbots and teachers show significant similarities whilst evaluating different types of essays. However, as aforementioned chatbots are slightly better at detecting the errors, mistakes, and misuse of grammar and vocabulary. This might be due to the fact that chatbots are incredibly mechanical and the functions of the algorithm make it easier for chatbots

to accomplish that. Additionally, teachers are better at recognising abstract concepts, this is a natural result of the human thinking skills that AI technologies are yet to develop. However, the findings showed slight differences in teacher feedback, it was rare, but it was there. The interview with the teachers points out that this might have occurred because of several factors. First of all, humans are organic beings, and they can get easily exhausted or tired, eventually affecting the way that teachers' grading and feedback. Secondly, personal bias or opinions might have a positive or negative effect on the teachers' feedback. Thirdly, teachers' perceptions of some criteria might differ, and this would affect the way that they give feedback. Last but not least, teachers might consider different factors as more important or vital for some types of essays eventually altering the way that they look at the essays.

On a similar note, not all chatbots were consistent as well. Out of three chatbots, ChatGPT proved to be the best option to use while detailly evaluating essays. Although they were mostly similar in grading, Gemini and BingAI showed repetition in feedback. Even though the repetition was not seen all the time, it was encountered enough times to have an effect on the results. Additionally, ChatGPT was more detailed and closer to the teachers in terms of analysing even abstract skills. It is not to say the other chatbots were bad, but ChatGPT is significantly more capable than the others at this point of time. It is worth mentioning that chatbots are as good as the used prompts. Prompts are critical when using chatbots, they are the guide for chatbots to follow. Since chatbots are computer programs, they literally follow the prompts exactly as given. That is why it is important to provide great prompts to chatbots. Adjusting the level of detail, explanation, suggestion, and the way the prompts are given would without a doubt significantly change the results. Additionally, although chatbots are pretty useful, they might not always be "perfect". Though rarely, chatbots might not function properly, therefore if teachers were to use chatbots as assistive tools, it is important to double-check the chatbots' responses.

Additionally, for feedback the data indicates that especially ChatGPT is quite successful and a considerable number of steps ahead of other chatbots with its significantly detailed, lengthy, and on point feedback. However, other chatbots are still somewhat effective to cover more general parts of the essay in their feedback. Teachers, in this equation, is in between ChatGPT and other chatbots, they can give

precise and concise feedback however they did not use exemplifying as much as ChatGPT. On a positive note, teachers had slightly more insight on the essays and conveyed more abstract concepts faintly better than ChatGPT and considerably better than other chatbots. In conclusion, the findings of this study suggest that especially the teachers might benefit from chatbots in terms of grading and giving quality feedback to essays with careful and considerate usage. Chatbots are proven to be efficient tools for evaluation, and they showed considerable similarities to teachers, even performed slightly better in some areas. All in all, in the age of technology and artificial intelligence, chatbots are invaluable assets that can prove useful in many areas including evaluation and assessment in ELT.

The findings of the study suggest that although chatbots perform decent, they might not be ready to replace teachers in terms of grading and giving feedback, but this study also showed that especially ChatGPT can be a significantly important assistive or helper tool to teachers whilst dealing with essays. Subsequently, chatbots are successful at many things but there is still a room for improvement for them to be perfect. However, teachers should benefit from chatbots as assistive tools to ease the burden off their shoulders. Chatbots are fast, quick, and accurate especially in analysing structural elements. Teachers can use hybrid evaluation techniques to quickly achieve better and more accurate results.

Needless to say, future ELT programs and teacher training should slightly adjust to the everlasting impact of AI and immense speed in which the AI technologies are improving. CALL has been a useful course for the field of ELT in Türkiye. However, in most universities, CALL courses show outdated resources to future English language teachers. Even the contemporary ones are yet to integrate AI technologies to classrooms and lessons. Teacher trainers, researchers, academicians, and universities should act quickly to include AI technologies in curricula. In-service teachers, pre-service teachers, students, and language learners should be familiar with AI technologies, their limitations, ethical aspects of using AI, and how best to use them. Moreover, AI enhanced modals can be used in every step of the way. Hybrid modelling can be embraced for almost every ELT course. Lastly, ELT researchers and AI developers should collaborate to create AI enhanced chatbots, and programs specific to the field of ELT even specific to evaluation and assessment in ELT.

4.1 Grading of the Essays

As numerous mentioned, AI enhanced multifunctional language generators also known as chatbots have been around for only 2 years, moreover they are getting better with more data, so they are everchanging. Chatbots are significantly different than the time they were first published, and it has not been long. Therefore, there is a lack of research revolving around chatbots, chatbots in ELT, chatbots in evaluation in ELT, and chatbots' grading. However, a variety of AI modals and technologies have been used in various ELT fields including grading and giving quality feedback. The findings of this study will be cross-referenced to both recent studies that includes recent enhanced chatbots and relatively older ones that uses AI technologies.

A study by Tulasi and Rao (2023) has found that integrating AI technologies into classroom and encouraging learners to use AI technologies are effective and critical to improve student learning and achievement. They further claimed that ELT and technology cannot be separated any longer because they are intertwined. This study by Tulasi and Rao (2023) highlighted the effectiveness of AI integration as was suggested in this particular research.

Even in the 2000 a study was conducted by Powers et al. (2000) that aimed compare automated essay scoring and human essay scoring. Even though this study was carried out 25 years ago, the findings revealed that there were modest similarities between teachers and automated graders. However, the similarities or relations were significantly weaker when compared to the present. Giving the nature of technology and the dramatic improvements in AI over the last few years this is an expected result. Additionally, Zupanc and Bosnic (2015) approached to the same topic but this time they had more advanced tools. According to their study, AI systems lacked in providing meaningful and detailed quality feedback to students, however they were successful in recognising certain errors, and giving automated feedback to correct those errors. Additionally, they were decent at grading therefore the researchers suggested that automated evaluation systems can serve as a complementary tool to teachers in terms of grading and giving feedback. Gierl et al. (2014) found that the agreement between computer classification and human raters were between 94.6% and 98.2%, however they highlighted that these systems may not be appropriate for all writing types. Another comprehensive study by Ifenthaler (2022) discussed the features of automated scoring systems and

concluded that although there were limitations such as bias, comprehensiveness and validity; automated scoring systems were at a point where they could provide insight into an individual's learning. He also suggested there were much to be done to mitigate the limitations of AI grading systems. Moving on to this year, Gobrecht et al. (2024) revealed findings that highly in accordance with this study. The team trained an open-source transformer model for grading. The findings of the study revealed that AI's error rate was smaller than teachers and AI proved to be more consistent. They suggested that AI can "...reduce human subjectivity, improve consistency and thus ultimately increase fairness." (Gobrecht et al., 2024) Finally, Floden (2024) compared the performance of ChatGPT in terms of grading university exams to human teachers. The results showed that ChatGPT gave higher scores than teachers. It was better at general questions but struggled to correctly score questions that are related to course lectures. It was revealed that ChatGPT can give acceptable scores that are close to human teachers. Moreover, Floden (2024) mentioned that teachers were surprised how well ChatGPT graded the questions. However, he highlighted some concerns such as ethical challenges and potential bias in training data.

Currently, there are no studies that compare chatbots within each other in terms of grading. The results of this study will significantly aid the literature in that aspect providing valuable results by comparing ChatGPT, BingAI, and Gemini. According to the findings of this study, significant differences have been found between ChatGPT, and BingAI and Gemini, whereas no significant differences were found between BingAI and Gemini in grading three different types of essays.

In conclusion, similar studies have been done for more than 20 years. It is clear that over the years AI technologies have improved marginally. Although for years the concerns remained similar, the biggest problem was AI technologies were yet to improve more. Although AI technologies and chatbots in particular in a better place now than even a few months ago, there is still a slight way to go to completely replace teachers in terms of grading. Additionally, almost all research has suggested that AI technologies should be used as assistive tools similar to this research. It can be said that this study shows the significant improvement of AI chatbots in terms of grading when considering other studies. Lastly, this study also shows the importance of using the correct AI tool, or chatbot in this case, to get the most efficient results.

4.2 Feedback for the Essays

Feedback too is not at a different position than grading. Although apart from grading, providing feedback is significantly different. Feedback reflects on the errors, mistakes, coulds, and shoulds of a written text or a speech. Therefore, it is more difficult to create systems that can analyse data to provide accurate descriptions and suggestions to learners and students. Consequently, it is critical to reflect on the previous work to especially understand the improvement of technology and AI enhanced feedback.

In accordance with the findings of this study, C. -J. Huang et al. (2008) published a study on the effectiveness of immediate feedback by e-learning tools. The team used an intelligent learning diagnosis tool that is based on text mining and machine learning techniques. The results indicated that the middle and low learners benefited from this tool to improve themselves on programming. Another study that is closer to this field was conducted by Chodorow et al. (2010) that aimed to evaluate two AWE systems. Two different systems provided feedback to two different groups. In one group, native and non-native speakers were given feedback on their article errors in essays. The other group benefited from feedback in writing emails. The results of the study showed that users were able to use valid corrections and made use of the given feedback effectively. One recent study by Teng (2024) focused on students' perceptions about the feedback from ChatGPT. The findings showed that AI had a significantly positive effect in assisting students to improve their writing skills as well as significantly increasing writing motivation, self-efficacy, engagement, and collaborative writing tendency. Teng's study also aligned with previous studies that previously demonstrated the facilitative role of AI in writing performance (Shen and Teng, 2024). Additionally, the integration of ChatGPT marginally enhanced teachers' self-efficacy and learners' writing skills.

Similar to grading, there are no studies that compare chatbots in terms of feedback. The findings of this study indicated interesting results regarding the feedback provided by different chatbots. Accordingly, the themes and mechanisms of feedback provided by ChatGPT, BingAI, and Gemini is almost identical. However, in terms of in-depth approaches such as exemplifying and suggestions, more detailed explanations, and critical comments for further improvement, ChatGPT is found to be more effective tool out of three.

In conclusion, AI and chatbots in feedback face some criticism especially in terms of ethics, bias, and the accuracy of nuanced feedback. However, consistent with previous studies, this study showed that, while not perfect, chatbots can now provide more in-depth feedback than before. Studies solely revolving around chatbots and feedback is very limited, yet existing studies show promising results. Moreover, this study will significantly contribute to the topic of quality feedback from chatbots with the data provided.

4.3 Conclusion

All things considered, this study is important to fill some gaps in this field considering the abilities of chatbots in terms of grading and providing quality feedback. There are some limitations especially concerning the sample sizes, and other limitations such as limited essay types, rubric and current abilities of chatbots follow the sample size. However, the provided data is invaluable and efficiently covers a considerable lack of research in ELT.

For further studies, researchers should prioritize increasing the sample sizes when trying to compare human teachers and chatbots. The perceptions of teachers, teacher educators, and future teachers in ELT are critical to understand the current place of AI in English language education, and ELT programs. Moreover, AI developers can collaborate with ELT researchers to create better and more efficient software to better respond to the needs of teachers and teacher educators. Lastly, it is important to integrate AI technologies and chatbots in teacher education programs to increase efficiency and AI literacy.

To sum up, this research revealed that chatbots are promising assets for teachers and learners in terms of grading and providing quality feedback for essays. Although there are minor differences between the gradings and feedback of AI and teachers, or AI and AI they do not pose a threat to cast a shadow on the feedback from AI. Moreover, there are some places where AI has performed slightly better and more consistent than teachers. Consequently, the conquest of AI is inevitable and teachers, teacher educators, and researchers should embrace these technologies and benefit from them as much as possible. Moreover, students can easily use chatbots to improve their writing skills with accurate grading and immediate feedback. Lastly, teachers should consider chatbots as assistive tools to help them in many tasks including grading and providing feedback to students.

5. CONCLUSION

In this section, implications, limitations of this study will be explored in detail. Furthermore, suggestions for future research will be made regarding the analysis of the data and the results of the study.

5.1 Implications of the Study

The use of Artificial Intelligence technologies has become a common strategy in almost every field of sciences and even more, our daily lives have started to become entangled with AI technologies. Arguably it is the biggest technological advancement in the history since the introduction of the internet. Humanity is at a point where the Nobel Prize in Physics is given to John Hopfield and Geoffrey Hinton, the pioneers of AI technologies, for using tools to develop methods that serve as the foundation of today's machine learning. Clearly, both natural sciences and social sciences significantly benefit from AI technologies. Moreover, English Language Teaching is a field that can successfully be aided by recent AI developments, in specific chatbots. Chatbots are extremely advanced language modals that effectively mimics human interaction both in written form and freshly spoken form. The field of ELT primarily focuses on language and humane interaction. Therefore, the possibilities that chatbots provide are almost uncountable.

In regard to that, this study aimed to further investigate the possibilities by comparing teachers and chatbots in addition to comparing chatbots within themselves in terms of grading and giving quality feedback. The analysis strongly suggests that chatbots are effective tools for this particular task. Considerably, additional steps can be taken to further improve the field and this technology.

One of the very first things that should be done would be to accept that AI is quickly invading everywhere and adjust the education and the system to effectively address the needs of modern technology. Literacy among fields is known to be critical, then there is media literacy followed by social media literacy, hereafter there needs to be strong emphasis on AI literacy. Students should see the capabilities of AI technologies and must be taught ethical aspects of this unlimited tool. Also, the prospect teachers should be aware of AI and consider that their students might use AI for certain tasks, hence they need to learn how to detect the

use of AI besides learning how to use it. Furthermore, teachers at all levels from K-12 to academy level should familiarize themselves with the ubiquitous nature of AI.

Additionally, for language teachers, teacher education stands at a critical point. Chatbots are now able to do many tasks that teachers would find time consuming such as administrative works, creating and evaluating exams, and even making short- and long-term curricula, lesson plans, and materials in correspondence. CALL is an important branch in ELT and AI should at least be as important as CALL and the necessary lessons should be provided to ELT students. As well as students, in-service teachers, lecturers, instructors, and even professors too should educate themselves in the subject. Moreover, the academic weight of the subject should be considered, and more academic studies need to be done to further improve the field and discover AI's possibilities that can aid ELT students, teachers, and academicians.

However, with all things considered AI literacy plays a critical role to ensure the validity and reliability of these technologies and chatbots. Users of these technologies of all levels must constantly double-check the results of their prompts. At least at this point no matter how capable they are, AI is nowhere near perfect, and it will take considerable time to smooth out the imperfections.

Another point would be the quest of integrating or merging the two together. It is suggested by literature and this study that AI chatbots is able to provide quick, accurate, and quality feedback. Yet, they are not able to comprehend the complexity and the depth of human thought processes. Teachers are more effective in terms of empathizing with the students and could understand what they want to mean. Therefore, a hybrid assessment modal can be proposed to enhance the quality of feedback and grading. AI would better focus on the mechanical aspects and teachers would provide deeper, and content-oriented feedback. This would not only help learners significantly better but would considerably decrease the time that teachers spend on evaluating essays.

Consequently, AI developers and AI researchers in the field of ELT can cooperate to create more comprehensive tools that can evaluate essays on more aspects. With enough data and the abilities of NLP and machine learning technologies could allow AI to assess even discourse-level features such as critical thinking and creativity.

In conclusion, the study indicates that AI and teachers can cooperate to complement each other that is until there are AI technologies that can delve into the complex minds of students, learners, and humans. Furthermore, a hybrid evaluation modal would enhance the quality of evaluation and assessment in addition to helping teachers by reducing their burden. Therefore, a possible curriculum might integrate AI-human mechanisms together to increase efficiency in terms of assessment procedures. Additionally, teacher education programs in ELT can include courses on how to implement AI tools into their teaching practice. Moreover, in-service teachers may need similar training, and this would allow teachers to optimise their time and offer more personalised guidance where it is most needed. All things considered, AI technologies offer nearly limitless opportunities and most of them are proven to be efficient, hence the teachers, and researchers would better shift their focuses on AI technologies to enhance their educational programs for all levels.

5.2 Limitations of the Study

Although the data being respectively reliable, like all studies this study has limitations as well. In a study like this, limitations can be classified considering human factor and AI factor. Both groups have different limitations that is worth exploring. Needless to say, that there are other limitations independent of those factors which will be explained in detail.

To begin with, one of the very first limitations that can be seen is the size of this particular study. The number of participants on both writing the essays and evaluating the essays could be a lot higher, resulting in more varied but significantly more reliable end outcomes. Moreover, the number of written essays by one person could be more than three with different and more specific tasks to get a greater number of written data. Besides the number of participants, competency of the participants could be a lot more different and wide-ranged. In this study, only a certain group of evaluators was used. Increasing or decreasing or doing both simultaneously could have observable effect on the outcome of the analysis, hence slightly adjusting the final figure. In addition, a major limitation is the lack of training for both students and evaluators. In this study, videos for participants were prepared on how to write an essay and different types of essays, and how to evaluate essays using TOEFL IBT writing rubric and each criterion were explained in detail. However, it is uncertain whether all participants watched the videos and understood

them without difficulty. Consequently, lack of training for teachers and students can be considered as another limitation. Lastly, as a major side effect of being “human”, reliability and consistency of the evaluators are of course a limitation. Human evaluations can be affected by far too many factors; psychology of the evaluator, beliefs of the evaluator, or even exhaustion of the evaluator are just a few of those factors. Even the evaluation of one essay can differ depending on those elements. In conclusion, although the researcher tried to limit the effects of limitations by human-evaluators, there are still considerable constraints that can be better adjusted for similar further studies.

Another element of limitations is on the side of the computers, more specifically AI-enhanced chatbots. Publication of chatbots is seen as one of the greatest breakthroughs in the history of technology. They changed almost everything, and they altered the way technology is used. However, since they are relatively novel technologies there is still a long way before they become perfect. One of the most important limitations of chatbots is the unreliability of chatbots. Although it is improving every second, sometimes chatbots may provide drastically different answers for same questions. Luckily, in this study chatbots were tested recurrently and the results were almost the same. Although it should be noted that this may not be the case all the time. Another limitation is chatbots are not omniscient. Their answers might be wrong, and they should not be trusted totally at this point of time. Users need AI literacy to analyse the data they get from AI sources. One considerable limitation is the prompts. Prompts are the way that shape the answers of chatbots. Different prompts lead to different results even when asking the same thing. That is why it is critical to find foolproof, complete prompts for chatbots the function the best. The last limitation is that chatbots are easily altered. This means that with enough data provided or with correct prompts users can change the way chatbots answer and even program them to give biased answers or political comments. However, all of the limitations of chatbots will possibly be overcome in the near future. Yet, users and researchers should be extra careful, critically analyse the data, and double-check the answers of the chatbots.

The last limitation that is worth mentioning are the other elements such as essays and rubric. It goes without saying that essays are a crucial component of this study. However, this study focuses on some specific types of essays which are narrative, expository, and argumentative. In order to get clearer and more precise

results, both types and numbers of essays can be increased. Moreover, there could be essays that focus on the same topic to better understand differences in writing. Even further, different types of writing could be explored to significantly enhance the scope of further studies. Another vital component is of course the rubric. TOEFL IBT writing rubric, which was accredited by numerous studies was used in this study. However, there were some concerns about the range of the rubric by teachers. The rubric is a wholistic rubric which may be deficient in feedback especially for some considerable criteria such as fairness. This study has utilised the rubric analytically which may be regarded as another limitation. For a better and drastically more detailed results, different rubrics can be used, improved or even created. Researchers can create study-specific rubrics or even essay specific rubric. Consequently, the results would naturally be affected when different rubrics were used.

5.3 Suggestions for Future Research

Considering the explored limitation there are few areas that can be suggested to researchers who want to further explore this particular topic in future studies. The most important aspect is the scale of the study, researchers would better investigate the results with a larger group of teachers and students alike. Different teachers at different levels of expertise and teach a variety of proficiency groups can be gathered for future studies to get more valid data. Furthermore, significantly a greater number of students can be included to write more essays to present a more diverse essay pool. Moreover, students with varying proficiency levels can be included to create drastically varied groups. Additionally, different essay types can be included to better test the limitations and capabilities of chatbots compared to teachers.

Another aspect would be to enhance the competencies of chatbots. In the future, researchers can collaborate with AI developers to create chatbots or tools that specialises in grading and giving feedback to essays. Furthermore, these tools can be further enhanced to accomplish a variety of evaluation and assessment tasks. For instance, these possible tools can be used to evaluate field specific courses. Additionally, prompt engineering can be made to get the most optimal results. With more improvement it is needless to say that chatbots would be even more efficient assets than they are now. Additionally, to complement each other, hybrid modelled

studies can be used that unite the chatbots and teachers to comprehend the collective abilities of the two groups.

Lastly, rubrics are critical to provide quality feedback and accurate grading. This study benefited from TOEFL IBT Writing Evaluation Rubric that is repeatedly proven valid and reliable. However, since it is a general rubric, it does not focus on the specific aspects of different essay types. To mitigate the effects of generalised rubric, more comprehensive and specific rubrics can be used for different essay types and competency levels.

In conclusion, all studies have relative limitations. For this study limitations were mainly based around three factors, the variety, the scale, and the methodology. Considering these factors and specific limitations; future researchers can include more participants in every step of their studies, use variety of methods, utilise different evaluation strategies, and alter the roles of chatbots and teachers to improve the evidence on the capabilities of chatbots in terms of grading and feedback and even more. This study and the following studies would play critical roles in determining the position of AI technologies in ELT.

5.4 Conclusion

All in all, this study aimed to investigate the possible differences and similarities between AI enhanced chatbots and teachers, and additionally, this study also examined the potential differences between ChatGPT, BingAI, and Gemini in terms of grading and giving feedback to narrative, expository, and argumentative essays that were written by first-grade ELT students study in Bolu Abant İzzet Baysal University. In order to successfully achieve this, a mixed-method approach was utilised. Initially, 10 students were asked to write three different types of essays. After the essays were collected, teachers and chatbots graded and gave feedback using TOEFL IBT Writing Evaluation rubric which is actually a holistic rubric, but this study benefited from it in an analytic way as well. This rubric explores the essays in four criteria that are task response, organisation, coherence, and language use also for further exploration overall grading/feedback. Accordingly, the quantitative data of the study consisted of the gradings that were made by teachers and chatbots. Each criterion was graded from 1 to 5 and the sum of the four criterion forms the overall grading. Furthermore, the qualitative data of the study includes the feedback provided by chatbots and teachers using the same rubric. Lastly, the gathered and analysed data of teachers and chatbots are compared

to get a glimpse of the capabilities of chatbots. Additionally, to get a better understanding of the thought processes of teachers, and their personal views on the evaluation process, a semi-structured interview was conducted with the teachers. These questions aimed to get the answers to how and why teachers do certain things whilst evaluating essays.

The findings of the quantitative data of the study indicate considerable similarities between chatbots and teachers whilst grading essays. According to the data, the normality tests indicated that teachers and ChatGPT had normal distribution for all essay types. On the other hand, BingAI had only one normal distribution and Gemini had zero normal distribution. Considering the normality tests, comparison tests were conducted. The findings of the comparison tests indicated that the teachers and ChatGPT had no significant differences in grading for all essay types. However, the results found significant differences between the teachers and BingAI in argumentative essays. Similarly, the teachers and Gemini had significant differences in expository essays. Additionally, for chatbots, ChatGPT and BingAI had significant differences in expository essays and argumentative essays, ChatGPT and Gemini had significant differences in argumentative essays, whereas BingAI and Gemini had no significant differences at all.

Followingly the analysis of the qualitative data showed similar results with some twists. First of all, the teachers and chatbots proved to be almost equal in terms of providing quality feedback. There were both advantages and disadvantages to either group, but in terms of highlighting the errors and considering certain themes, both groups performed similarly. As for differences, teachers had more variety of themes and sub-themes which indicated that teachers consider more factors when dealing with essays, these factors are mostly abstract and more nuanced concepts that chatbots are yet to conquer. However, chatbots performed better in terms of providing concise suggestions and exemplify their feedback using passages, words, phrases, and more from the essays in addition to that, chatbots were slightly more successful whilst analysing the mechanical aspects of the language. Followingly, the differences between chatbots indicated that ChatGPT was more effective than BingAI and Gemini with more advanced feedback supported by examples, more suggestions, and detailed explanations. All in all, both

group of evaluators did certain aspects more than the other which supporting a possible collaboration between the two.

Overall, both for grading and giving quality feedback teachers and chatbots performed similarly with both groups having specific advantages. Consequently, the study suggests that chatbots can be effective assistive tools to teachers aiding them to take some burden out of them whilst evaluating essays. Furthermore, it is within learners' reach to efficiently use these tools to their advantages. Therefore, there needs to be more research regarding the usefulness of AI. Additionally, teacher education courses and education system should benefit from the results of this study and further studies. Furthermore, teachers and prospective teachers in ELT should educate themselves on the possibilities of AI. Lastly, universities should include AI as a lesson or part of a lesson in their ELT programmes. All in all, AI is now almost everywhere and seems to be occupying more parts of humans' lives every day. Subsequently, the power of AI should be embraced and put the use of advantages of ELT teachers, researchers, academicians, and students alike. With this way more things would be significantly easier, and teachers would focus on teaching more efficiently.

REFERENCES

- Ai, H. (2017). Providing graduated corrective feedback in an intelligent computer-assisted language learning environment. *ReCALL*, 29(3), 313–334. <https://doi.org/10.1017/s095834401700012x>
- Akayoğlu, S. (2011). An analysis of text based cmc of advanced efl learners in second life. [Doctoral Dissertation, Middle East Technical University]. Ulusal Tez Merkezi.
- Akayoğlu, S. (2019). Theoretical frameworks used in CALL studies: A systematic review. *Teaching English with Technology*, 19(4), 104–118.
- Albirini, A. (2006). Teachers' attitudes toward information and communication technologies: The case of syrian efl teachers. *Computers & Education*, 47(4), 373–398. <https://doi.org/10.1016/j.compedu.2004.10.013>
- Alharbi, W. (2023). Ai in the foreign language classroom: A pedagogical overview of automated writing assistance tools. *Education Research International*, 2023, 1–15. <https://doi.org/10.1155/2023/4253331>
- Ali, Z. (2020). Artificial Intelligence (AI): A Review of its uses in language teaching and learning. *IOP Conference Series: Materials Science and Engineering*, 769(1), 1–6. <https://doi.org/10.1088/1757-899x/769/1/012043>
- Alkureishi, M. A., Choo, Z.-Y., Rahman, A., Ho, K., Benning-Shorb, J., Lenti, G., Velázquez Sánchez, I., Zhu, M., Shah, S. D., & Lee, W. W. (2021). Digitally disconnected: Qualitative study of patient perspectives on the digital divide and potential solutions. *JMIR Human Factors*, 8(4), e33364. <https://doi.org/10.2196/33364>
- Allenby, B. R., & Sarewitz, D. (2011). The techno-human condition. *The MIT Press*. <https://doi.org/10.7551/mitpress/8714.001.0001>
- Allwright, R., & Bailey, K. M. (1991). *Focus on the language classroom*. In Howard B. A., & Peter S. (Eds) Cambridge: Cambridge University Press.
- Anderson, S. J. (1976, October). TICCIT Project. Paper presented at the National Association of Educational Broadcasters, Chicago, Illinois. ERIC Document No. ED134226. Retrieved from www.eric.ed.gov/
- Anh, D. T., & Marginson, S. (2013). Global learning through the lens of Vygotskian sociocultural theory. *Critical Studies in Education*, 54(2), 143–159. <https://doi.org/10.1080/17508487.2012.722557>
- Applebee, A. N., & Langer, J. A. (2011). EJ Extra: A snapshot of writing instruction in middle schools and high schools. *English Journal*, 100(6), 14–27. <https://doi.org/10.58680/ej201116413>
- Asad, M. M., Naz, A., Churi, P., & Tahanzadeh, M. M. (2021). Virtual reality as pedagogical tool to enhance experiential learning: A systematic literature review. *Education Research International*, 2021, 1–17. <https://doi.org/10.1155/2021/7061623>
- Ashfaq, M. W., Tharewal, S., Iqbal, S., & Kayte, C. N. (2020). A review on techniques, characteristics and approaches of an intelligent tutoring chatbot system. In *Proceedings of the 2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC)*, 258–262. <https://doi.org/10.1109/icsidempc49020.2020.9299583>

- Atkins, N.E. & Vasu, E.S. (2000). Measuring knowledge of technology usage and stages of concern about computing: A study of middle school teachers. *Journal of Technology and Teacher Education*, 8(4), 279-302.
- Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research*, 13(2), 111–127. <https://doi.org/10.2190/9lmd-3u28-3a0g-ftqt>
- Bai, L., & Hu, G. (2016a). In the face of fallible awe feedback: How do students respond? *Educational Psychology*, 37(1), 67–81. <https://doi.org/10.1080/01443410.2016.1223275>
- Bakker, A. (2004). In *Design research in statistics education: On symbolizing and Computer Tools*. essay, CD-β Press, Center for Science and Mathematics Education.
- Bakker, A., & Gravemeijer, K. P. (2006). An historical phenomenology of mean and Median. *Educational Studies in Mathematics*, 62(2), 149–168. <https://doi.org/10.1007/s10649-006-7099-8>
- Bakker, A., & Hoffmann, M. H. (2005). Diagrammatic reasoning as the basis for developing concepts: A semiotic analysis of Students' learning about statistical distribution. *Educational Studies in Mathematics*, 60(3), 333–358. <https://doi.org/10.1007/s10649-005-5536-8>
- Bancheri, S. (2006). A language teacher's perspective on effective courseware. In Randall P. D., & Margaret A. H. (Eds.) *Changing Language Education through CALL* (pp. 31–47). New York: Routledge.
- Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education*, 21(1), 1–15. <https://doi.org/10.1186/s41239-024-00455-4>
- Barab, S. A., & Luehmann, A. L. (2003). Building sustainable science curriculum: Acknowledging and accommodating local adaptation. *Science Education*, 87(4), 454–467. <https://doi.org/10.1002/sce.10083>
- Bassein, R., & Underwood, J. (1985). *Juegos Comunicativos: Games for communicative practice in Spanish, based on Puntos de Partida*. New York: Random House, Inc.
- Bax, S. (2003). Call—past, present and future. *System*, 31(1), 13–28. [https://doi.org/10.1016/s0346-251x\(02\)00071-4](https://doi.org/10.1016/s0346-251x(02)00071-4)
- Beatty, K. (2015). *Teaching and researching computer-assisted language learning*. Routledge.
- Bebell, D., & O'Dwyer, L. (2010). Educational Outcomes and Research from 1:1 Computing Settings. *The Journal of Technology, Learning and Assessment*, 9(1), 5-15
- Bellman, R. (1978). *An introduction to artificial intelligence: Can computers think?* Boyd & Fraser Pub. Co.
- Biber, D., Nekrasova, T., & Horn, B. (2011). The effectiveness of feedback for L1-English and l2-writing development: A meta-analysis. *ETS Research Report Series*, 2011(1), i–99. <https://doi.org/10.1002/j.2333-8504.2011.tb02241.x>
- Bitchener, J., & Knoch, U. (2009). The contribution of written corrective feedback to language development: A ten month investigation. *Applied Linguistics*, 31(2), 193–214. <https://doi.org/10.1093/applin/amp016>

- Bitzer, D. L. (1986). The PLATO project at the university of Illinois. *Engineering Education*, 77(3), 175–80.
- Boden, M. A. (Ed.). (1996). *Artificial intelligence*. Elsevier.
- Boon, P. (2009). A designer speaks: Designing educational software for 3D geometry. *Educational Designer*, 1(2). <http://www.educationaldesigner.org/ed/volume1/issue2/article7> (Accessed 18 January 2025).
- Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education*, 38(6), 698–712. <https://doi.org/10.1080/02602938.2012.691462>
- Brandt, D. (2005). Writing for a living: Literacy and the knowledge economy. *Written Communication*, 22(2), 166–197. <https://doi.org/10.1177/0741088305275218>
- Brown, H. D. (1993). *Principles of language learning and teaching*. In H. Douglas Brown. Prentice Hall Regents.
- Brown, H. D. (2007). *Teaching by Principle; An interactive approach to language pedagogy*. Third Edition. New York: Pearson Longman.
- Buchanan, J., Purcell, K., & Friedrich, L. (2013). *The impact of digital tools on student writing and how writing is taught in schools*. National Writing Project Pew Research Center.
- Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The criterion online writing service. *Ai magazine*, 25(3), 27–27. <https://doi.org/10.1002/rcm.5057>.
- Cambridge University Press. (n.d.). Artificial. In [dictionary.cambridge.org](https://dictionary.cambridge.org/dictionary/english/artificial). Retrieved March 9, 2024, from <https://dictionary.cambridge.org/dictionary/english/artificial>
- Cambridge University Press. (n.d.). Intelligence. In [dictionary.cambridge.org](https://dictionary.cambridge.org/dictionary/english/intelligence). Retrieved March 9, 2024, from <https://dictionary.cambridge.org/dictionary/english/intelligence>
- Campbell-Howes, K. (2021, April 28). *Why is ai a good thing for language teachers and learners?* Oxford TEFL. Teacher Training. <https://oxfordtefl.com/blog/why-is-ai-a-good-thing-for-language-teachers-and-learners/>
- Carey, M. (2004). Call visual feedback for pronunciation of vowels. *CALICO Journal*, 21(3), 571–601. <https://doi.org/10.1558/cj.v21i3.571-601>
- Celce-Murcia, M. (2001). *Teaching English as a second or foreign language*. Third Edition. Heinle: Language Learning.
- Chapelle, C. A. (2007). *Computer applications in second language acquisition*. Cambridge University Press.
- Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing*, 32(3), 385–405. <https://doi.org/10.1177/0265532214565386>
- Charniak, E., & McDermott, D. (1985). *Introduction to artificial intelligence*. Addison-Wesley.
- Chaudron, C. (1977). A descriptive model of discourse in the corrective treatment of learners' errors. *Language Learning*, 27(1), 29–46. <https://doi.org/10.1111/j.1467-1770.1977.tb00290.x>

- Chen, C. F., & Cheng, W. Y. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology*, 12(2), 94-112.
- Chen, C. H., Koong, C.-S., & Liao, C. (2022). Influences of integrating dynamic assessment into a speech recognition learning design to support pupils' English speaking skills, learning anxiety and cognitive load. *Educational Technology & Society*, 25(1), 1-14.
- Chen, M. X., Lee, B. N., Bansal, G., Cao, Y., Zhang, S., Lu, J., & Wu, Y. (2019). *Gmail smart compose: Real-time assisted writing*. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2287-2295).
- Chen, Y.-L., Hsu, C.-C., Lin, C.-Y., & Hsu, H.-H. (2022). Robot-assisted language learning: Integrating Artificial Intelligence and virtual reality into English tour guide practice. *Education Sciences*, 12(7), 437. <https://doi.org/10.3390/educsci12070437>
- 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>
- Chon, Y. V., Shin, D., & Kim, G. E. (2021). Comparing L2 learners' writing against parallel machine-translated texts: Raters' assessment, linguistic complexity and errors. *System*, 96, 102408. <https://doi.org/10.1016/j.system.2020.102408>
- Chun, D., Kern, R., & Smith, B. (2016). Technology in language use, language teaching, and language learning. *The Modern Language Journal*, 100(S1), 64-80. <https://doi.org/10.1111/modl.12302>
- Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education*. Routledge.
- Coley, R. J., Cradler, J., & Engel, P. K. (1997). Computers and classrooms: The status of technology in U.S. schools. *ETS Policy Information Report*, 1-69.
- Cooke-Plagwitz, J. (2013). New Directions in call: An objective introduction to Second Life. *CALICO Journal*, 25(3), 547-557. <https://doi.org/10.1558/cj.v25i3.547-557>
- 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>
- Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of chatgpt. *Innovations in Education and Teaching International*, 61(2), 228-239. <https://doi.org/10.1080/14703297.2023.2190148>
- Courts, B., & Tucker, J. (2012). Using technology to create a dynamic classroom experience. *Journal of College Teaching & Learning (TLC)*, 9(2), 121-128. <https://doi.org/10.19030/tlc.v9i2.6907>
- Creswell, J. W. (2004). Designing a mixed methods study in primary care. *The Annals of Family Medicine*, 2(1), 7-12. <https://doi.org/10.1370/afm.104>
- Crompton, H., Edmett, A., Ichaporia, N., & Burke, D. (2024). AI and English language teaching: Affordances and challenges. *British Journal of Educational Technology*, 55(6), 2503-2529. <https://doi.org/10.1111/bjet.13460>
- Cummings Hlas, A., Conroy, K., & Hildebrandt, S. A. (2017). Student teachers and call: Personal and pedagogical uses and beliefs. *CALICO Journal*, 34(3), 336-354. <https://doi.org/10.1558/cj.26968>

- Çakmak, F. (2022). Chatbot-human interaction and its effects on EFL pupils' L2 speaking performance and anxiety. *Novitas- ROYAL (Research on Youth and Language)*, 16(2), 113–131.
- Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y. S., Gašević, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. *IEEE International Conference on Advanced Learning Technologies (ICALT)* (pp. 323–325).
- Dai, W., Tsai, Y.-S., Fan, Y., Gašević, D., & Chen, G. (2022). Measuring inconsistency in written feedback: A case study in politeness. *Lecture Notes in Computer Science*, 560–566. https://doi.org/10.1007/978-3-031-11644-5_49
- Davies, G., & Steel, D. (1981, April). First steps in computer-assisted language learning at Ealing College of Higher Education. Paper presented at the *CALL 81 Symposium*, University of Leeds. Retrieved from www.ict4lt.org/en/Davies_Steel_1981.doc
- Davies, G., & Higgins, J. (1982). *Using computers in Language learning: A teacher's guide*. Centre for Information on Language Teaching and Research.
- Davies, G., Otto, S. E., & Rüschhoff, B. (2013). Historical perspectives on CALL. *Contemporary computer-assisted language learning*, 19–38.
- Devi, V. A., & Revathi, K. (2016). A brief review of literature on immediate feedback studies in call. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 8(3), 262–268. <https://doi.org/10.21659/rupkatha.v8n3.27>
- Dewey, J. (1892). *Psychology*. Harper&Brothers. 1–6.
- Dikli, S., & Bleyle, S. (2014). Automated Essay Scoring Feedback for Second Language Writers: How Does it compare to instructor feedback? *Assessing Writing*, 22, 1–17. <https://doi.org/10.1016/j.asw.2014.03.006>
- Dixon, S. (2015). The pastoral potential of audio feedback: A review of the literature. *Pastoral Care in Education*, 33(2), 96–104. <https://doi.org/10.1080/02643944.2015.1035317>
- Dizon, G., & Gayed, J. M. (2021). Examining the impact of grammarly on the quality of mobile L2 writing. *The JALT CALL Journal*, 17(2), 74–92. <https://doi.org/10.29140/jaltcall.v17n2.336>
- Dizon, G., & Tang, D. (2020). Intelligent personal assistants for autonomous second language learning: An investigation of Alexa. *The JALT CALL Journal*, 16(2), 107–120. <https://doi.org/10.29140/jaltcall.v16n2.273>
- Doughty, C. (2001). Cognitive underpinnings of focus on form. In P. Robinson (Ed.), *Cognition and second language instruction*. pp. 206–57. Cambridge: Cambridge University Press.
- Drijvers, P. (2015). Digital technology in mathematics education: Why it works (or doesn't). In *Selected regular lectures from the 12th international congress on mathematical education*. pp. 135-151. Springer International Publishing.
- Duhon, G. J., House, S., Hastings, K., Poncy, B., & Solomon, B. (2014). Adding immediate feedback to explicit timing: An option for enhancing treatment intensity to improve mathematics fluency. *Journal of Behavioral Education*, 24(1), 74–87. <https://doi.org/10.1007/s10864-014-9203-y>
- Eden, C. A., Chisom, O. N., & Adeniyi, I. S. (2024). Harnessing technology integration in education: Strategies for enhancing learning outcomes and equity. *World Journal of Advanced*

- Egbert, J., Paulus, T., & Nakamichi, Y. (2002). The impact of CALL instruction on classroom computer use: A foundation for rethinking technology in teacher education. *Language Learning & Technology*, 6(3), 108–126.
- Elliot, N., A. Ruggles Gere, G. Gibson, C. Toth, C. Whithaus, and A. Presswood. (2013). Use and limitations of automated writing evaluation software. *WPA-CompPile Research Bibliographies* 23.
- Ellis, R. (1997). *SLA research and language teaching*. Cambridge University Press.
- Elola, I., & Oskoz, A. (2016). Supporting second language writing using multimodal feedback. *Foreign Language Annals*, 49(1), 58–74. <https://doi.org/10.1111/flan.12183>
- Englert, C. S., Mariage, T. V., & Dunsmore, K. (2006). Tenets of sociocultural theory in writing instruction research. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 208- 221). Guilford.
- Evans, K. C., & Teichman, J. (1991). *Philosophy: A beginner's guide Jenny Teichman & Katherine C. Evans*. Blackwell Publishers.
- Fajfar, P., Campitelli, G., & Labollita, M. (2012). Effects of immediacy of feedback on estimations and performance. *Australian Journal of Psychology*, 64, 167–177. <https://doi.org/10.1111/j.0004-9530.2011.00048.x>
- Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of CHATGPT: Implications for educational practice and Research. *Innovations in Education and Teaching International*, 61(3), 460–474. <https://doi.org/10.1080/14703297.2023.2195846>
- Feng, H.-H., Saricaoglu, A., & Chukharev-Hudilainen, E. (2016). Automated error detection for developing grammar proficiency of ESL learners. *CALICO Journal*, 33(1), 49–70. <https://doi.org/10.1558/cj.v33i1.26507>
- Ferreira, A., Moore, J. D., & Mellish, C. (2007). A study of feedback strategies in foreign language classrooms and tutorials with implications for intelligent computer-assisted language learning systems. *International Journal of Artificial Intelligence in Education*, 17(4), 389–422.
- Fetzer, J. H. (1990). What is Artificial Intelligence? *Studies in Cognitive Systems*, 3–27. https://doi.org/10.1007/978-94-009-1900-6_1
- Firat, E. A., & Firat, S. (2020). Web 3.0 in learning environments: A systematic review. *Turkish Online Journal of Distance Education*, 22(1), 148–169.
- Firth, A. (1996). The discursive accomplishment of normality: On ‘lingua franca’ English and conversation analysis. *Journal of Pragmatics*, 26(2), 237–259. [https://doi.org/10.1016/0378-2166\(96\)00014-8](https://doi.org/10.1016/0378-2166(96)00014-8)
- Fitria, T. N. (2024). Using chatbot-based artificial intelligence (AI) for writing an English essay: The ability of CHATGPT, perplexity AI, and Chatsonic. *Journal of Language Intelligence and Culture*, 6(2), 103–128. <https://doi.org/10.35719/jlic.v6i2.139>

- 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*, 1–24. <https://doi.org/10.1002/berj.4069>
- Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z. (2017). Stimulating and sustaining interest in a language course: An experimental comparison of chatbot and human task partners. *Computers in Human Behavior*, 75, 461–468. <https://doi.org/10.1016/j.chb.2017.05.045>
- Fryer, L. K., Nakao, K., & Thompson, A. (2019). Chatbot learning partners: Connecting learning experiences, interest and competence. *Computers in Human Behavior*, 93, 279–289. <https://doi.org/10.1016/j.chb.2018.12.023>
- Fryer, L., Coniam, D., Carpenter, R., & Lăpușneanu, D. (2020). Bots for language learning now: Current and future directions. *Language Learning and Technology*, 24(2), 8–22.
- Fung, W. J. (2010). A predictive text completion software in Python. *The Python Paper s Monograph*, 2.
- Gallacher, A., Thompson, A., & Howarth, M. (2018). “My robot is an idiot!” – students’ perceptions of AI in the L2 classroom. *Future-Proof CALL: Language Learning as Exploration and Encounters – Short Papers from EUROCALL 2018*, 70–76. <https://doi.org/10.14705/rpnet.2018.26.815>
- Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by CHATGPT to real abstracts with detectors and blinded human reviewers. *Npj Digital Medicine*, 6(1), 1–5. <https://doi.org/10.1038/s41746-023-00819-6>
- Gierl, M. J., Latifi, S., Lai, H., Boulais, A.-P., & De Champlain, A. (2014). Automated essay scoring and the future of educational assessment in medical education. *Medical Education*, 48(10), 950–962. <https://doi.org/10.1111/medu.12517>
- Gobrecht, A., Tuma, F., Möller, M., Zöller, T., Zakhvatkin, M., Wuttig, A., & Schütt, S. (2024). Beyond human subjectivity and error: a novel AI grading system. *arXiv preprint arXiv:2405.04323*.
- Goody, J. (1986). *The logic of writing and the Organization of Society*. Cambridge University Press.
- Grabe, W., & Kaplan, R. B. (2014). *Theory and Practice of Writing*. <https://doi.org/10.4324/9781315835853>
- Graham, S. (2019). Changing how writing is taught. *Review of Research in Education*, 43(1), 277–303. <https://doi.org/10.3102/0091732x18821125>
- Graham, S., Harris, K. R., & Hebert, M. A. (2011). *Informing writing: The benefits of formative assessment*. Carnegie Corporation of New York.
- Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A Meta Analysis. *The Elementary School Journal*, 115(4), 523–547. <https://doi.org/10.1086/681947>
- Grant Jun Otsuki Lecturer in Cultural Anthropology. (2024, December 10). *OK computer: To prevent students cheating with AI text-generators, we should bring them into the classroom*. The Conversation. <https://theconversation.com/ok-computer-to-prevent-students-cheating-with-ai-text-generators-we-should-bring-them-into-the-classroom-129905>

- 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). 1–44
- Hamouma, C., & Menezla, N. (2019). The impact of digital literacy proficiency on EFL students' academic writing performance. *International Journal of Digital Literacy and Digital Competence*, 10(4), 40–55. <https://doi.org/10.4018/ijdlldc.2019100103>
- Haristiani, N. (2019). Artificial Intelligence (AI) chatbot as language learning medium: An inquiry. *Journal of Physics: Conference Series*, 1387(1), 012020. <https://doi.org/10.1088/1742-6596/1387/1/012020>
- Hart, R. S. (1995). The Illinois PLATO foreign languages project. In R. Sanders (Ed.) (1995), Thirty years of computer-assisted language instruction: Festschrift for John R. Russell, *CALICO Journal*, 12(4), 15–37
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Havranek, G., & Cesnik, H. (2001). Factors affecting the success of corrective feedback. *EUROSLA Yearbook*, 1, 99–122. <https://doi.org/10.1075/eurosla.1.10hav>
- He, Y. (2021). Challenges and countermeasures of translation teaching in the era of Artificial Intelligence. *Journal of Physics: Conference Series*, 1881(2), 022086. <https://doi.org/10.1088/1742-6596/1881/2/022086>
- Heift, T. (2004). Corrective feedback and learner uptake in call. *ReCALL*, 16(2), 416–431. <https://doi.org/10.1017/s0958344004001120>
- Herrington, A., and C. Moran. (2012). 'Writing to a machine is not writing at all' In N. Elliot and L. Perelman (eds.). *Writing Assessment in the Twenty First Century: Essays in Honor of Edward M. White*. New York: Hampton Press, pp. 219–32.
- Hew, K. F., Huang, W., Du, J., & Jia, C. (2022). Using chatbots to support student goal setting and social presence in fully online activities: Learner Engagement and perceptions. *Journal of Computing in Higher Education*, 35(1), 40–68. <https://doi.org/10.1007/s12528-022-09338-x>
- Higgins, J., & Johns, T. (1984). *Computers in language learning*. Collins ELT; Addison-Wesley, World Language Division.
- Hinkel, E. (2005). What research on second language writing tells us and what it doesn't. In *Handbook of Research in Second language teaching and learning* (pp. 523–538). L. Erlbaum Associates; Routledge.
- Hirvela, A. (2005). Computer-based reading and writing across the curriculum: Two case studies of L2 writers. *Computers and Composition*, 22(3), 337–356. <https://doi.org/10.1016/j.compcom.2005.05.005>
- Hockly, N. (2018). Automated writing evaluation. *ELT Journal*, 73(1), 82–88. <https://doi.org/10.1093/elt/ccy044>
- Hong, K. H. (2010). Call teacher education as an impetus for L2 teachers in integrating technology. *ReCALL*, 22(1), 53–69. <https://doi.org/10.1017/s095834400999019x>
- Hope, G. R., Taylor, H. F., & Pusack, J. P. (1984). *Using computers in teaching foreign languages*. Harcourt, Brace, Jovanovich.

- Hoven, D. (2007). The affordances of technology for student teachers to shape their teacher education experience. In Kassen, M. A. et al. (Eds.), *Preparing and developing technology-proficient L2 teachers* (pp. 133–163). San Marcos, TX: CALICO
- Howatt, A. P.R. (1997). Talking shop: Transformation and change in ELT. *ELT Journal*, 51(3), 263–268. <https://doi.org/10.1093/elt/51.3.263>
- Howatt, Anthony Philip Reid, & Widdowson, H. G. (1984). *A history of English language teaching*. Oxford University Press.
- Huang, C.-J., Chen, C.-H., Luo, Y.-C., Chen, H.-X., & Chuang, Y.-T. (2008). Developing an intelligent diagnosis and assessment eLearning tool for introductory programming. *Educational Technology & Society*, 11(4), 139–157 <https://www.jstor.org/stable/jeductechsoci.11.4.139>.
- Huang, W., Hew, K. F., & Fryer, L. K. (2021). Chatbots for language learning—are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, 38(1), 237–257. <https://doi.org/10.1111/jcal.12610>
- Hubbard, P. (2004). Learner training for effective use of CALL. In S. Fotos & C. Browne (Eds.), *New perspectives on CALL for second language classrooms* (pp. 45–68). Mahwah, NJ: Lawrence Erlbaum.
- Hubbard, P. (2008). CALL and the future of language teacher education. *CALICO Journal*, 25(2), 175–188. Hubbard, P., & Levy, M. (2006). Introduction. In Hubbard, P. & Levy, M. (Eds.), *Teacher Education and CALL* (pp. ix–xi). Philadelphia: John Benjamins. <https://doi.org/10.1075/llt.14>
- Hubbard, P., & Levy, M. (2006). The scope of CALL education. In Hubbard, P. & Levy, M. (Eds.), *Teacher Education and CALL* (pp. 3–20). Philadelphia: John Benjamins. <https://doi.org/10.1075/llt.14.04hub>
- Hung, S.-T. A. (2016). Enhancing feedback provision through Multimodal Video Technology. *Computers & Education*, 98, 90–101. <https://doi.org/10.1016/j.compedu.2016.03.009>
- Hunt, E. B. (2014). *Artificial intelligence*. Academic Press.
- Hunter, D., & Smith, R. (2012). Unpackaging the past: ‘CLT’ through ELTJ keywords. *ELT Journal*, 66(4), 430–439. <https://doi.org/10.1093/elt/ccs036>
- Hussein, K. O. F., Phelps, L., & Bessent, H. (1980). *CALIS*. Durham, NC: Duke University.
- Hyland, K. (2016). *Second language writing*. Cambridge University Press.
- Hyland, K., & Hyland, F. (2019). *Feedback in second language writing: Contexts and issues*. Cambridge University Press.
- Ibrahim, K. (2023). Using AI-based detectors to control AI-assisted plagiarism in ESL writing: “The terminator versus the machines.” *Language Testing in Asia*, 13(1), 48. <https://doi.org/10.1186/s40468-023-00260-2>
- Ifenthaler, D. (2022). Automated essay scoring systems. In O. Zawacki-Richter & I. Jung (Eds.), *Handbook of open, distance and digital education* (pp. 1–15). Springer. https://doi.org/10.1007/978-981-19-0351-9_59-1

- Ilies, R., Judge, T. A., & Wagner, D. T. (2010). The influence of cognitive and affective reactions to feedback on subsequent goals. *European Psychologist*, 15(2), 121–131. <https://doi.org/10.1027/1016-9040/a000011>
- Ima, N., & Jihad, A. (2024). Computer-assisted language learning: The impact in language education. *Vifada Journal of Education*, 2(1), 36–42. <https://doi.org/10.70184/smkf7m80>
- Ippolito, D., Yuan, A., Coenen, A., & Burnam, S. (2022). Creative writing with an ai-powered writing assistant: Perspectives from professional writers. *arXiv preprint arXiv:2211.05030*.
- Ira Irzawati, Agnes Felisya Unamo, Agnes Agnes, & Vinny Angelika. (2024). The use of chat GPT in ELT. *International Journal of Education, Language, Literature, Arts, Culture, and Social Humanities*, 2(1), 32–44. <https://doi.org/10.59024/ijellacush.v2i1.674>
- Ivankova, N. V., & Creswell, J. W. (2009). Mixed methods. *Qualitative research in applied linguistics: A practical introduction*, 23, 135–161.
- Jackson, P. C. (2019). *Introduction to artificial intelligence*. Courier Dover Publications.
- Jafarian, K., Soori, A., & Kafipour, R. (2012). The effect of computer assisted language learning (CALL) on EFL high school students' writing achievement. *European Journal of Social Sciences*, 27(2), 138–148.
- Jiang, H., Cheng, Y., Yang, J., & Gao, S. (2022). AI-powered chatbot communication with customers: Dialogic interactions, satisfaction, engagement, and customer behavior. *Computers in Human Behavior*, 134, 107329. <https://doi.org/10.1016/j.chb.2022.107329>
- Jin, L., & Cortazzi, M. (2011). Re-evaluating traditional approaches in Second Language Teaching and Learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2) (pp. 558–575). Routledge. ISBN 978-0-415-99872-7.
- Jones, R. L. (1995). TICCIT and CLIPS: The early years. In R. Sanders (Ed.), *Thirty years of computer-assisted language instruction: Festschrift for John R. Russell*, *CALICO Journal*, 12(4), 84–96.
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130–144. <https://doi.org/10.1016/j.edurev.2007.05.002>
- Juzwik, M. M., Curcic, S., Wolbers, K., Moxley, K. D., Dimling, L. M., & Shankland, R. K. (2006). Writing into the 21st century. *Written Communication*, 23(4), 451–476. <https://doi.org/10.1177/0741088306291619>
- Kang, E., & Han, Z. (2015). The efficacy of written corrective feedback in improving L2 written accuracy: A meta-analysis. *The Modern Language Journal*, 99(1), 1–18. <https://doi.org/10.1111/modl.12189>
- Kannan, J., & Munday, P. (2018). New trends in second language learning and teaching through the lens of ICT, networked learning, and artificial intelligence. *Círculo de Lingüística Aplicada a La Comunicación*, 76, 13–30. <https://doi.org/10.5209/clac.62495>
- Kargar, N., & Tayebipour, F. (2015). The effect of scaffolding on EFL learners' reading comprehension. *Modern Journal of Language Teaching Methods*, 5(4), 446–453.
- Kazu, İ. Y., & Kuvvetli, M. (2023). The influence of pronunciation education via artificial intelligence technology on vocabulary acquisition in learning English. *International Journal*

of Psychology and Educational Studies, 10(2), 480–493.
<https://doi.org/10.52380/ijpes.2023.10.2.1044>

Kenning, M. J., & Kenning, M.-M. (1984). *Introduction to computer assisted language teaching* M. J. Kenning; M.-M. Kenning. Oxford University Press.

Kerly, A., Hall, P., & Bull, S. (2007). Bringing chatbots into education: Towards natural language negotiation of open learner models. *Knowledge-Based Systems*, 20(2), 177–185.
<https://doi.org/10.1016/j.knosys.2006.11.014>

Kessler, G. (2006). Assessing CALL teacher training: What are we doing and what could we do better? In P. Hubbard & M. Levy (Eds.), *Teacher education in CALL* (pp. 23–44). Philadelphia

Kessler, G. (2007). Formal and informal call preparation and teacher attitude toward technology. *Computer Assisted Language Learning*, 20(2), 173–188.
<https://doi.org/10.1080/09588220701331394>

Kessler, G. (2010). When they talk about call: Discourse in a required call class. *CALICO Journal*, 27(2), 376–392. <https://doi.org/10.11139/cj.27.2.376-392>

Kessler, G., & Plakans, L. (2001). Incorporating ESOL learners' feedback and usability testing in instructor-developed call materials. *TESOL Journal*, 10(1), 15–20.
<https://doi.org/10.1002/j.1949-3533.2001.tb00012.x>

Khalifa, M., & Albadawy, M. (2024). Using artificial intelligence in academic writing and research: An essential productivity tool. *Computer Methods and Programs in Biomedicine Update*, 5, 100145. <https://doi.org/10.1016/j.cmpbup.2024.100145>

Khalil, M., & Er, E. (2023). Will CHATGPT get you caught? Rethinking of plagiarism detection. *Lecture Notes in Computer Science*, 475–487. https://doi.org/10.1007/978-3-031-34411-4_32

Khoii, R., & Doroudian, A. (2013). Automated Scoring of EFL Learners' Written Performance: a Torture or a Blessing? In Conference proceedings. *ICT for language learning* (p. 367).

Kim, H., Shin, D. K., Yang, H., & Lee, J. H. (2019). A study of AI chatbot as an assistant tool for school English curriculum. *Korean Association For Learner-Centered Curriculum And Instruction*, 19(1), 89–110. <https://doi.org/10.22251/jlcci.2019.19.1.89>

Kim, H., Yang, H., Shin, D., & Lee, J. H. (2022). Design principles and architecture of a second language learning chatbot. *Language Learning & Technology*, 26(1), 1–18.

Kim, M. C., & Hannafin, M. J. (2004). Designing online learning environments to support scientific inquiry. *Quarterly Review of Distance Education*, 5(1), 1–10.

Kim, M. C., Hannafin, M. J., & Bryan, L. A. (2007). Technology-Enhanced inquiry tools in science education: An emerging pedagogical framework for classroom practice. *Science Education*, 91(6), 1010–1030. <https://doi.org/10.1002/sce.20219>

Kim, N. Y. (2018). A study on chatbots for developing Korean college students' English listening and reading skills. *Journal of Digital Convergence* 16(8) 19–26.

Kim, N. Y. (2019). A study on the use of artificial intelligence chatbots for improving English grammar skills. *Journal of Digital Convergence*, 17(8).

- Kim, N.-Y. (2022). AI-Integrated Mobile-Assisted Language Learning: Is it an effective way of preparing for the TOEIC test in classroom environments? *English Teaching*, 77(3), 79–102. <https://doi.org/10.15858/engtea.77.3.202209.79>
- Kirschenbaum, M. (2017). *Track changes: A literary history of word processing*. Cambridge, MA and London, England: Harvard University Press. <https://doi.org/10.4159/9780674969469>
- Kiuhara, S. A., Graham, S., & Hawken, L. S. (2009). Teaching writing to high school students: A national survey. *Journal of Educational Psychology*, 101(1), 136–160. <https://doi.org/10.1037/a0013097>
- Kıymaz, M. (2022). Pedagogically sound: Audio feedback as an alternative to written feedback in distance education: EFL/ESL students' and teachers' perceptions and its effectiveness. *English Language Education*, 81–105. https://doi.org/10.1007/978-3-031-13540-8_5
- Klein, S., Aeschlimann, J. F., Balsiger, D. F., Converse, S. L., Court, C., Foster, M., Lao, R., Oakley, J. D., & Smith, J. (1979). Automatic novel writing: A status report. *Text Processing / Textverarbeitung*, 338–412. <https://doi.org/10.1515/9783110837537-015>
- Kohnke, L. (2022a). A pedagogical chatbot: A supplemental language learning tool. *RELC Journal*, 54(3), 828–838. <https://doi.org/10.1177/00336882211067054>
- Kohnke, L. (2022b). A qualitative exploration of student perspectives of chatbot use during emergency remote teaching. *International Journal of Mobile Learning and Organisation*, 16(4), 475. <https://doi.org/10.1504/ijmlo.2022.125966>
- Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *Relc Journal*, 54(2), 537–550.
- Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by grammarly: A multiple case study. *Assessing Writing*, 44, 100450. <https://doi.org/10.1016/j.asw.2020.100450>
- Kostka, I., & Toncelli, R. (2023). Exploring applications of chatgpt to English language teaching: Opportunities, challenges, and recommendations. *Teaching English as a Second or Foreign Language--TESL-EJ*, 27(3), n3. <https://doi.org/10.55593/ej.27107int>
- Kracht, M. (2007). *Introduction to linguistics*. Los Angeles. LA: Hilgard Avenue, 2.
- Kreindler, I. (1998). Designing feedback that is hard to ignore: A boost from multimedia. In K. Cameron (Ed.). *Multimedia CALL: Theory and practice*. pp. 243–50).
- Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2022). Interacting with educational Chatbots: A systematic review. *Education and Information Technologies*, 28(1), 973–1018. <https://doi.org/10.1007/s10639-022-11177-3>
- Kumaravadivelu, B. (1994). The postmethod condition: (e)merging strategies for second/foreign language teaching. *TESOL Quarterly*, 28(1), 27–48. <https://doi.org/10.2307/3587197>
- Kumaravadivelu, B. (2012). Individual identity, cultural globalization, and teaching English as an international language: the case for an epistemic break. In L. Alsagotf, S. L. McKay, G. Hu, and W. A. Renandya (eds.), *Principles and practices for teaching English as an international language*. New York: Routledge. 9–27.
- Kurzweil, R., & Jaroach, D. (1990). *The age of Intelligent Machines*. MIT Press.

- Lam, Y. (2000). Technophilia vs. Technophobia: A preliminary look at why second-language teachers do or do not use technology in their classrooms. *The Canadian Modern Language Review*, 56(3), 389–420. <https://doi.org/10.3138/cmlr.56.3.389>
- Larsen-Freeman, D., and Anderson, M. (2011). *Techniques & principles in language teaching*. Oxford: Oxford University Press.
- Lavolette, E., Polio, C., & Kahng, J. (2015). The accuracy of computer-assisted feedback and students' responses to it. *Language Learning & Technology*, 19(2), 50–68.
- Lee, D., Kim, H., & Sung, S.-H. (2022). Development research on an AI English learning support system to facilitate learner-generated-context-based learning. *Educational Technology Research and Development*, 71(2), 629–666. <https://doi.org/10.1007/s11423-022-10172-2>
- Lee, J. H., Yang, H., Shin, D., & Kim, H. (2020). Chatbots. *ELT Journal*, 74(3), 338–344. <https://doi.org/10.1093/elt/ccaa035>
- Levy, M. (1996). *Call: Context and conceptualisation*. Oxford University Press.
- Levy, Mike, & Stockwell, G. (2006). *Call dimensions options and issues in computer-assisted language learning*. Routledge.
- Li, J., & Cumming, A. (2001). Word processing and second language writing: A longitudinal case study. *International Journal of English Studies*, 1(2), 127–152.
- Li, Jiang. (2006). The mediation of technology in ESL writing and its implications for writing assessment. *Assessing Writing*, 11(1), 5–21. <https://doi.org/10.1016/j.asw.2005.09.001>
- Li, Jinrong, Link, S., & Hegelheimer, V. (2015). Rethinking the role of Automated Writing Evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing*, 27, 1–18. <https://doi.org/10.1016/j.jslw.2014.10.004>
- Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System*, 44, 66–78. <https://doi.org/10.1016/j.system.2014.02.007>
- Liao, H.-C. (2015). Using automated writing evaluation to reduce grammar errors in writing. *ELT Journal*, 70(3), 308–319. <https://doi.org/10.1093/elt/ccv058>
- Liao, H.-C. (2016). Enhancing the grammatical accuracy of EFL writing by using an awe-assisted process approach. *System*, 62, 77–92. <https://doi.org/10.1016/j.system.2016.02.007>
- Linn, M. (2003). Technology and science education: Starting points, research programs, and Trends. *International Journal of Science Education*, 25(6), 727–758. <https://doi.org/10.1080/09500690305017>
- Liou, H.-C. (2013). Can grammatical call help EFL writing instruction? *CALICO Journal*, 10(1), 23–44. <https://doi.org/10.1558/cj.v10i1.23-44>
- Little, D. G. (1989). *Self-access systems for language learning*. Authentik Language Learning Resources Ltd.
- Liu, S.-C., & Hung, P.-Y. (2016). Teaching pronunciation with computer assisted pronunciation instruction in a technological university. *Universal Journal of Educational Research*, 4(9), 1939–1943. <https://doi.org/10.13189/ujer.2016.040902>

- Liu, Y., Theodore, P., & Lavelle, E. (2004). Experimental effects of online instruction on teachers' concerns about technology integration. *International Journal of Instructional Technology and Distance Learning*, 1(1), Retrieved from: http://www.itdl.org/journal/Jan_04/article03.htm
- Liutkus, D. (2012). Investigating call in the classroom: Situational variables to consider. *Advances in Language and Literary Studies*, 3(1), 107–118. <https://doi.org/10.7575/aiac.alls.v.3n.1p.107>
- Livingstone, K. A. (2012). The Importance of feedback and reinforcement in computer-assisted language learning (CALL). *Baraton Interdisciplinary Research Journal*, 2(1), 43–51.
- Lo, S. (2023). Neural machine translation in EFL classrooms: Learners' vocabulary improvement, immediate vocabulary retention and delayed vocabulary retention. *Computer Assisted Language Learning*, 1–20. <https://doi.org/10.1080/09588221.2023.2207603>
- Lock, J., & Redmond, P. (2021). Embedded experts in online collaborative learning: A case study. *The Internet and Higher Education*, 48, 100773. <https://doi.org/10.1016/j.iheduc.2020.100773>
- Lord, G., & Lomicka, L. (2011). Calling on educators: Paving the way for the future of technology and CALL. In Arnold, N. & Ducate, L. (Eds.), *Present and future promises of CALL: From theory and research to new directions in language teaching* (pp. 441–469). San Marcos, TX: CALICO.
- Luke, C., & Britten, J. (2013). Expanding role of technology in foreign language teacher education programs. *CALICO Journal*, 24(2), 253–268. <https://doi.org/10.1558/cj.v24i2.253-268>
- Lund, B., & Ting, W. (2023). Chatting about chatgpt: How may AI and GPT impact academia and libraries? *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4333415>
- Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing*, 18(1), 30–43. <https://doi.org/10.1016/j.jslw.2008.06.002>
- Lyons, J. (1968). *Introduction to theoretical linguistics*. Cambridge University Press.
- Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake. *Studies in Second Language Acquisition*, 19(1), 37–66. <https://doi.org/10.1017/s0272263197001034>
- Mabuan, R. A. (2024). Chatgpt and ELT: Exploring teachers' voices. *International Journal of Technology in Education*, 7(1), 128–153. <https://doi.org/10.46328/ijte.523>
- Masuda, K., & Arnett, C. (2015). 1. cognitive linguistics, sociocultural theory and language teaching: Introduction. *Cognitive Linguistics and Sociocultural Theory*, 1–22. <https://doi.org/10.1515/9781614514442-003>
- McCarthy, J. (2020). Student perceptions of screencast video feedback for summative assessment tasks in the creative arts. *Advances in Higher Education and Professional Development*, 177–192. <https://doi.org/10.4018/978-1-7998-0426-0.ch009>
- McKee, H. A., & Porter, J. E. (2020). Ethics for AI writing. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 110–116. <https://doi.org/10.1145/3375627.3375811>
- Meo, S. (2013). Giving feedback in medical teaching: A case of lung function laboratory/spirometry. *College of Physicians & Surgeons Pakistan*, 23(1), 86–89.
- Mizumoto, A., & Eguchi, M. (2023). *Exploring the Potential of Using an AI Language Model for Automated Essay Scoring*. <https://doi.org/10.2139/ssrn.4373111>

- Mohamed, A. (2021). Feedback in computer-assisted language learning: A meta-analysis. *Proceedings of the 2021 AERA Annual Meeting*. <https://doi.org/10.3102/1681250>
- Moore, K. A., Rutherford, C., & Crawford, K. A. (2016). Supporting postsecondary english language learners' writing proficiency using Technological Tools. *Journal of International Students*, 6(4), 857–872. <https://doi.org/10.32674/jis.v6i4.321>
- Nagata, N. (1993). Intelligent computer feedback for second language instruction. *The Modern Language Journal*, 77(3), 330. <https://doi.org/10.2307/329102>
- Nagata, N. (1997). The effectiveness of computer-assisted Metalinguistic instruction: A case study in Japanese. *Foreign Language Annals*, 30(2), 187–200. <https://doi.org/10.1111/j.1944-9720.1997.tb02341.x>
- Nagata, N. (2013). Computer vs. workbook instruction in Second language acquisition. *CALICO Journal*, 14(1), 53–75. <https://doi.org/10.1558/cj.v14i1.53-75>
- Naik, U., & Shivalingaiah, D. (2008). Comparative study of web 1.0, web 2.0 and web 3.0. *International CALIBER*, 2, 28–29
- National Teacher Training Institute (NTTI). (2011). Thirteen ed online. Retrieved <http://www.thirteen.org/edonline/ntti/resources/video1.html>.
- Nattie Dehn. 1981. Story generation after TALE-SPIN. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, Vol. 1. 16–18.
- Nazari, N., Shabbir, M. S., & Setiawan, R. (2021). Application of artificial intelligence powered digital writing assistant in higher education: Randomized controlled trial. *Heliyon*, 7(5). <https://doi.org/10.1016/j.heliyon.2021.e07014>
- Neil McIntyre and Mirella Lapata. (2009). Learning to tell tales: A data-driven approach to story generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: (1). Association for Computational Linguistics*, 217–225.
- Nilsson, N. J. (1982). *Principles of Artificial Intelligence* Nils J. Nilsson. Springer.
- Nilsson, N. J. (2009). *The quest for artificial intelligence A history of ideas and achievements* Nils J. Nilsson. Cambridge Univ. Press.
- Nobles, S., & Paganucci, L. (2015). Do digital writing tools deliver? student perceptions of writing quality using digital tools and online writing environments. *Computers and Composition*, 38, 16–31. <https://doi.org/10.1016/j.compcom.2015.09.001>
- Olson, D. R. (2009). The history of writing. In *The Sage handbook of writing development* (pp. 6–16).
- OpenAI. (2023). Chat GPT. Retrieved from <https://openai.com/blog/chatgpt/> on 2 January 2023.
- Parra G., L., & Calero S., X. (2019). Automated writing evaluation tools in the improvement of the writing skill. *International Journal of Instruction*, 12(2), 209–226. <https://doi.org/10.29333/iji.2019.12214a>
- Peters, M. (2006). 9. developing computer competencies for pre-service language teachers. *Language Learning & Language Teaching*, 153–165. <https://doi.org/10.1075/llt.14.14pet>
- Plakans, L., Gebril, A., & Bilki, Z. (2019). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing*, 36(2), 161–179. <https://doi.org/10.1177/0265532216669537>

- Pokrivcakova, S. (2015). CALL and Foreign Language Education. <https://www.kajk.pf.ukf.sk/public/S.%20Pokrivčáková%20et%20al.%20-%20CALL%20and%20Foreign%20Language%20Education.pdf>.
- Pokrivcakova, S. (2019). Preparing teachers for the application of AI-powered technologies in foreign language education. *Journal of Language and Cultural Education*, 7(3), 135–153. <https://doi.org/10.2478/jolace-2019-0025>
- Poole, D. L., Mackworth, A. K., & Goebel, R. (1998). *Computational intelligence: A logical approach*. Oxford University Press.
- Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2000). Comparing the validity of automated and human essay scoring. *ETS Research Report Series*, 2000(2). <https://doi.org/10.1002/j.2333-8504.2000.tb01833.x>
- Radhamani, R., Sasidharakurup, H., Sujatha, G., Nair, B., Achuthan, K., & Diwakar, S. (2014). Virtual Labs improve student's performance in a classroom. *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, 138–146. https://doi.org/10.1007/978-3-319-13293-8_17
- Rakıcıoğlu-Söylemez, A., & Akayoğlu, S. (2019). Prospective EFL teachers' perceptions of using call in the classroom. *Advances in Higher Education and Professional Development*, 189–204. <https://doi.org/10.4018/978-1-5225-8583-1.ch011>
- Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2016). Automated writing evaluation for Formative Assessment of Second Language Writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology*, 37(1), 8–25. <https://doi.org/10.1080/01443410.2015.1136407>
- Rane, N., Choudhary, S., & Rane, J. (2023). Education 4.0 and 5.0: Integrating Artificial Intelligence (AI) for personalized and adaptive learning. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4638365>
- Richards, J. C. (1985). The secret life of methods. In J. C. Richards, *The context of language teaching*. New York: Cambridge University Press. 32–45.
- Richards, J. C., & Rodgers, T. S. (2014). *Approaches and methods in language teaching*. Cambridge University Press.
- Rilling, S., Dahlman, A., Dodson, S., Boyles, C., & Pazvant, O. (2013). Connecting call theory and practice in Preservice teacher education and beyond. *CALICO Journal*, 22(2), 213–235. <https://doi.org/10.1558/cj.v22i2.213-235>
- Robinson, G. L. (1985). *Computer-assisted-instruction in Foreign Language Education: A comparison of the effectiveness of different methodologies and different forms of error correction: Final report*. Center for Language & Crosscultural Skills; U.S. Dept. of Education, International Research and Studies Division.
- Roscoe, R. D., Wilson, J., Johnson, A. C., & Mayra, C. R. (2017). Presentation, expectations, and experience: Sources of student perceptions of automated writing evaluation. *Computers in Human Behavior*, 70, 207–221. <https://doi.org/10.1016/j.chb.2016.12.076>
- Roumeliotis, K. I., & Tselikas, N. D. (2023). Chatgpt and open-AI models: A preliminary review. *Future Internet*, 15(6), 192. <https://doi.org/10.3390/fi15060192>
- Rubin, J., Ediger, A., Coffin, E., Van Handle, D., & Whiskeyman, A. (1990). Survey of interactive videodiscs, *CALICO Journal*, 7(3), 31–56.
- Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence: A modern approach*. Prentice-Hall.

- Rüschhoff, B. (2002a). Authenticity and innovative learning scenarios in technology enriched environments for language learning. In T. Abbott et al. (Eds.), *Paradigms shift. Implications of emerging technologies for teaching languages & cultures*. Proceedings of Digital Stream 4. Monterey: California State University Monterey Bay.
- Rüschhoff, B. (2002b). Languages. In H. D. Adelsberger, B. Collins & J. M. Pawlowski (Eds.), *Handbook on information technologies for education and training* (pp. 523–43). Berlin: Springer Verlag.
- Rüschhoff, B. (2009). Output-oriented language learning with digital media. In M. Thomas (Ed.), *Handbook of research on Web 2.0 and second language learning* (pp. 42–59). Hershey, PA: Information Science Reference
- Salaberry, M. R. (2001). The use of technology for Second language learning and teaching: A retrospective. *The Modern Language Journal*, 85(1), 39–56. <https://doi.org/10.1111/0026-7902.00096>
- Sanders, R. H. (2013). Thirty years of computer assisted language instruction. *CALICO Journal*, 12(4), 6–14. <https://doi.org/10.1558/cj.v12i4.6-14>
- Sanders, R. H., & Sanders, A. F. (2013). History of an AI spy game. *CALICO Journal*, 12(4), 114–127. <https://doi.org/10.1558/cj.v12i4.114-127>
- Schneider, E. W., & Bennion, J. L. (1984). Veni, vidi, vici, via videodisc: A simulator for instructional courseware. In D. H. Wyatt (Ed.), *Computer-assisted language instruction* (pp. 41–6). Oxford: Pergamon.
- Schulze, M. (2013). Ai in call--artificially inflated or almost imminent? *CALICO Journal*, 25(3), 510–527. <https://doi.org/10.1558/cj.v25i3.510-527>
- Seedhouse, P. (1997). The case of the missing “no”: The relationship between pedagogy and interaction. *Language Learning*, 47(3), 547–583. <https://doi.org/10.1111/0023-8333.00019>
- Seidlhofer, B. (2005). English as a lingua franca. *ELT journal*, 59(4), 339–341
- Shermis, M. D., & Burstein, J. (2003). *Automated Essay Scoring: A cross-disciplinary perspective*. New Jersey.
- Shield, L. (2003). MOO as a language learning tool. In U. Felix (Ed.), *Language learning online: Towards best practice* (pp. 97–122). Lisse: Swets and Zeitlinger.
- Shim, Y. (2013). The effects of online writing evaluation program. *Teaching English with Technology*, 13(3), 18–34
- Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly*, 27(4), 657–677. <https://doi.org/10.2307/3587400>
- Slattery, P. J., & Kowalski, R. (1998). On screen: The composing processes of first-year and upper-level college students. *Computers and Composition*, 15(1), 61–81. [https://doi.org/10.1016/s8755-4615\(98\)90025-1](https://doi.org/10.1016/s8755-4615(98)90025-1)
- Smoker, B. L. and T., Hassinger, S., & Loukides, M. (n.d.). *Radar - O'Reilly*. O'Reilly Media. <http://radar.oreilly.com/2006/12/web-20-compact-definition-tryi.html>
- Spada, N., & Lightbown, P. M. (1993). Instruction and the development of questions in L2 classrooms. *Studies in Second Language Acquisition*, 15(2), 205–224. <https://doi.org/10.1017/s0272263100011967>

- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and CHATGPT feedback of students' writing. *Learning and Instruction*, 91, 101894. <https://doi.org/10.1016/j.learninstruc.2024.101894>
- Stephens, T. M., Ahmad, K., Corbett, G., Rogers, M., & Sussex, R. (1987). Computers, Language Learning and Language Teaching. *Hispania*, 70(1), 108. <https://doi.org/10.2307/343679>
- Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing*, 19, 51–65. <https://doi.org/10.1016/j.asw.2013.11.007>
- Straková, Z., & Cimermanová, I. (2018). Developing reflective skills of student teachers in the virtual learning environment. *Electronic Journal of e-Learning*, 16(2), 107-121.
- Susnjak, T., & McIntosh, T. (2024). CHATGPT: The end of online exam integrity? *Education Sciences*, 14(6), 656. <https://doi.org/10.3390/educsci14060656>
- Tebben, J. R. (1996). *A course in medical and technical terminology*. Burgess Pub. Co.
- Teng, M. F. (2024). "CHATGPT is the companion, not enemies": EFL learners' perceptions and experiences in using chatgpt for feedback in writing. *Computers and Education: Artificial Intelligence*, 7, 100270. <https://doi.org/10.1016/j.caeai.2024.100270>
- Thomas, Michael. (2009). *Handbook of Research on web 2.0 and second language learning*. Information Science Reference.
- Tokoz-Goktepe, F. (2014). Speaking problems of 9 th grade high school Turkish learners of L2 English and possible reasons for those problems: Exploring the teachers and students' perspectives. *Procedia - Social and Behavioral Sciences*, 116, 1875–1879. <https://doi.org/10.1016/j.sbspro.2014.01.487>
- Tulasi, L., & Rao, C. S. (2023). Integration of AI-Technologies into ELT: A brief study. *Journal for Research Scholars and Professionals of English Language Teaching*, 7(38), 1–7. <https://doi.org/10.54850/jrspelt.7.38.003>
- Turing, A. (1950). Computing Machinery and intelligence. *Mind*, LIX (236), 433–460. <https://doi.org/10.1093/mind/lix.236.433>
- Turmudi, D. (2017). Rethinking academic essay writing: Selected genres in comparison. *Premise: Journal of English Education*, 6(2), 119–138. <https://doi.org/10.24127/pj.v6i2.1052>
- Van Lier, L. (1996). *Interaction in the language curriculum: awareness, autonomy, and authenticity*. New York: Longman.
- Von Der Emde, S., Schneider, J., & Kötter, M. (2001). Technically speaking: Transforming language learning through virtual learning environments (moos). *The Modern Language Journal*, 85(2), 210–225. <https://doi.org/10.1111/0026-7902.00105>
- Wadhwa, D. (2015). A study of web 3.0 technologies for collaborative learning. *International Journal of Engineering Technology, Management and Applied Sciences*, 3, 92-97.
- Waldron, S., Wood, C., & Kemp, N. (2016). Use of predictive text in text messaging over the course of a year and its relationship with spelling, orthographic processing and grammar. *Journal of Research in Reading*, 40(4), 384–402. <https://doi.org/10.1111/1467-9817.12073>
- Wang, P. (2013). Can automated writing evaluation programs help students improve their English writing? *International Journal of Applied Linguistics & English Literature*, 2(1), 6–12. <https://doi.org/10.7575/ijalel.v.2n.1p.6>

- Wang, P. L. (2015). Effects of an automated writing evaluation program: Student experiences and perceptions. *Electronic Journal of Foreign Language Teaching*, 12(1).
- Wang, Y.-J., Shang, H.-F., & Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a Foreign Language University Students' writing. *Computer Assisted Language Learning*, 26(3), 234–257. <https://doi.org/10.1080/09588221.2012.655300>
- Ware, P. (2005). Automated writing evaluation as a pedagogical tool for writing assessment. In A. Pandian, G. Chakravarthy, P. Kell, & S. Kaur (Eds.), *Strategies and practices for improving learning and literacy* (pp. 174-184). Selangor, Malaysia: University Putra Malaysia Press.
- 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., & Healey, D. (1998). Computers and language learning: An overview. *Language Teaching*, 31(2), 57–71. <https://doi.org/10.1017/s0261444800012970>
- 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>
- Warschauer, M., Kern, R., & Kern, R. G. (2000). *Network-based language teaching: Concepts and practice*. Cambridge Univ. Press.
- Webster's New World Dictionary. (n.d.). philosophy
- Williams, L., Abraham, L. B., & Bostelmann, E. D. (2014). A discourse-based approach to call training and professional development. *Foreign Language Annals*, 47(4), 614–629. <https://doi.org/10.1111/flan.12119>
- 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., & MacArthur, C. (2024). Exploring the role of automated writing evaluation as a formative assessment tool supporting self-regulated learning in writing. *The Routledge International Handbook of Automated Essay Evaluation*, 197–220. <https://doi.org/10.4324/9781003397618-14>
- Wilson, J., & Roscoe, R. D. (2019). 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, J., Myers, M. C., & Potter, A. (2022). Investigating the promise of automated writing evaluation for supporting formative writing assessment at scale. *Assessment in Education: Principles, Policy & Practice*, 29(2), 183–199. <https://doi.org/10.1080/0969594x.2022.2025762>
- Winkler, R., & Soellner, M. (2018). Unleashing the potential of Chatbots in Education: A state-of-the-art analysis. *Academy of Management Proceedings*, 2018(1), 15903. <https://doi.org/10.5465/ambpp.2018.15903abstract>
- Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2016). Supporting learners' agentic engagement with feedback: a systematic review and a taxonomy of recipience processes. *Educational Psychologist*, 52(1), 17–37. <https://doi.org/10.1080/00461520.2016.1207538>
- Yang, N. (2004). Using My Access! in EFL writing. In *The proceedings of 2004 International Conference and Workshop on TEFL & Applied Linguistics* (pp. 550-564). Taipei, Taiwan: Ming Chuan University.

- Zaghlool, Z. D. (2020). The impact of using call online writing activities on EFL university students' writing achievement. *Theory and Practice in Language Studies*, 10(2), 141–148. <https://doi.org/10.17507/tpls.1002.01>
- Zaini, A., & Mazdayasna, G. (2014). The effect of computer assisted language learning on the development of EFL learners' writing skills. *Procedia - Social and Behavioral Sciences*, 98, 1975–1982. <https://doi.org/10.1016/j.sbspro.2014.03.631>
- Zeevy-Solovey, O. (2024). Comparing peer, CHATGPT, and teacher corrective feedback in EFL writing: Students' perceptions and preferences. *Technology in Language Teaching & Learning*, 6(3), 1482. <https://doi.org/10.29140/tltl.v6n3.1482>
- Zhang, D., Fan, Y., & Du, W. (2013). Sociocultural theory applied to Second language learning: Collaborative learning with reference to the Chinese context. *International Education Studies*, 6(9), 165–174. <https://doi.org/10.5539/ies.v6n9p165>
- Zhang, Z. (Victor), & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing*, 36, 90–102. <https://doi.org/10.1016/j.asw.2018.02.004>
- Zhang, Z. (Victor). (2016). Student engagement with computer-generated feedback: A case study. *ELT Journal*. <https://doi.org/10.1093/elt/ccw089>
- Zheng, D., Bischoff, M., & Gilliland, B. (2015). Vocabulary learning in massively multiplayer online games: Context and action before words. *Educational Technology Research and Development*, 63(5), 771–790. <https://doi.org/10.1007/s11423-015-9387-4>
- Zupanc, K., & Bosnić, Z. (2015). Advances in the field of automated essay evaluation. *Informatica*, 39(4), 383–395

APPENDICES

Appendix 1 Example of an Evaluation Sheet for Teachers

Essay Type	Narrative Essay 1
Title of the Essay	A DAY WHEN EVERYTHING WENT WRONG
Task Response	
Organisation	
Coherence	
Language Use (Mechanics)	
Overall Feedback	

Appendix 2 TOEFL IBT Writing Evaluation Rubric



TOEFL

TOEFL® iBT Test Independent Writing Rubrics (Scoring Standards)

Score	Task Description
5	An essay at this level largely accomplishes all of the following: - effectively addresses the topic and task - is well organized and well developed, using clearly appropriate explanations, exemplifications, and/or details - displays unity, progression, and coherence - displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors
4	An essay at this level largely accomplishes all of the following: - addresses the topic and task well, though some points may not be fully elaborated - is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications, and/or details - displays unity, progression, and coherence, though it may contain occasional redundancy, digression, or unclear connections - displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form, or use of idiomatic language that do not interfere with meaning
3	An essay at this level is marked by one or more of the following: - addresses the topic and task using somewhat developed explanations, exemplifications, and/or details - displays unity, progression, and coherence, though connection of ideas may be occasionally obscured - may demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning - may display accurate but limited range of syntactic structures and vocabulary
2	An essay at this level may reveal one or more of the following weaknesses: - limited development in response to the topic and task - inadequate organization or connection of ideas - inappropriate or insufficient exemplifications, explanations, or details to support or illustrate generalizations in response to the task - a noticeably inappropriate choice of words or word forms - an accumulation of errors in sentence structure and/or usage
1	An essay at this level is seriously flawed by one or more of the following weaknesses: - serious disorganization or underdevelopment - little or no detail, or irrelevant specifics, or questionable responsiveness to the task - serious and frequent errors in sentence structure or usage
0	An essay at this level merely copies words from the topic, rejects the topic, or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.

Copyright © 2008 by Educational Testing Service. All rights reserved.

Appendix 3 Tables of the Descriptive Analyses

Narrative Essays

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
NTRALL	5	3.90	4.30	4.0833	.13292
NORALL	5	3.50	4.00	3.6667	.17512
NCOALL	5	3.50	4.10	3.8167	.24833
NLUALL	5	2.90	3.80	3.4167	.32506
NOALL	5	14.40	15.50	14.9833	.35449
Valid N (listwise)	5				

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
TRALL	3	3.90	4.10	4.0000	.10000
ORALL	3	3.30	3.90	3.5667	.30551
COALL	3	3.20	3.70	3.4000	.26458
LUALL	3	2.60	3.20	2.8333	.32146
OVALL	3	13.40	14.20	13.8000	.40000
Valid N (listwise)	3				

Expository Essays

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
TRALL	5	3.80	4.50	4.1800	.27749
ORALL	5	3.40	4.30	3.8400	.40373
COALL	5	3.70	4.40	4.1800	.27749
LUALL	5	4.00	4.40	4.1800	.17889
OVALL	5	15.20	17.40	16.3800	.94446
Valid N (listwise)	5				

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
TRALL	3	3.80	4.15	3.9833	.17559
ORALL	3	3.40	4.15	3.7833	.37528
COALL	3	3.20	4.15	3.6667	.47522
LUALL	3	3.35	3.65	3.5000	.15000
OVALL	3	13.90	16.10	14.9333	1.10604
Valid N (listwise)	3				

Argumentative Essays

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
ATRALL	5	3.80	4.90	4.4200	.52631
AORALL	5	2.50	4.50	3.6300	.82432
ACOALL	5	3.50	4.70	4.0800	.47117
ALUALL	5	3.80	4.30	4.0200	.21679
AOVALL	5	15.10	18.40	16.3800	1.54822
Valid N (listwise)	5				

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
TRALL	3	3.55	4.00	3.8167	.23629
ORALL	3	3.45	4.00	3.8000	.30414
COALL	3	3.20	3.90	3.5500	.35000
LUALL	3	2.80	3.35	3.0500	.27839
OVALL	3	13.00	14.75	13.8833	.87512
Valid N (listwise)	3				

Appendix 4 Sample Essays and Feedback

Sample Narrative Essay

A Day of Unfortunate Events

It was a day like any other, or so I thought as I woke up to the gentle rays of the morning sun. Little did I know that it would become a day etched in my memory as the one when everything went wrong.

The troubles began innocently enough. As I rushed through my morning routine, I realized that I had overslept and was already running late for school. Hastily grabbing a coffee to go, I raced out the door, I sprained my foot. Frustration set in, but I managed to go to the school and attend the lesson. I couldn't listen to the teacher carefully because of the severe pain in my foot.

The day's misfortunes continued to unfold like a series of unfortunate events. After the lesson, I decided to go to the hospital. I went there alone because I did not want to worry my family. I waited for hours to get examined. I stood and waited and the doctor did not pay attention to me. Finally, the doctor wrote out a prescription and I left the hospital.

As if sensing my despair, the universe conspired to add more obstacles to my already challenging day. I went to the pharmacy and it was closed. So I had to find a pharmacy on duty. The pharmacy I found was far from the center and I had to walk a long way. Despite this, I walked and took the medicine prescribed by the doctor. Meanwhile, my foot started to hurt more.

With a heavy heart, I started to walk back to the bus stop. Unfortunately, the buses didn't stop to pick me up because they were full of people. I called a taxi. The taxi arrived quite late. The taxi driver charged me too much. It was very late when I got home. My family was worried a lot about me.

The climax of the day unfolded when I realised that I had forgotten my phone at the pharmacy. I couldn't shake the feeling that the universe had chosen this day to test the limits of my resilience. In the midst of this calamity, however, a silver lining emerged. The pharmacist found my home address and gave it to me. My family rallied around me, offering support and encouragement. Their words of comfort served as a reminder that, despite the chaos, I was not alone in facing life's challenges.

As the day drew to a close, I reflected on the myriad of setbacks and the unexpected display of kindness from some people. It became evident that even in the darkest moments, there is a glimmer of hope and a reminder that setbacks are merely temporary. The day when everything went wrong turned out to be a test of my resilience, teaching me to navigate adversity with a newfound strength. I was happy that I did not face a bigger calamity. I prayed for the things that I had. I thank god for giving me all I want and I wished.

Sample Teacher Feedback/Grading

Essay Type	Narrative Essay 7
Title of the Essay	A Day of Unfortunate Events
Task Response	The moment I began reading the very first sentence, I was suspicious of AI. If he is not a poet, it is impossible to write such a poetic essay. But here it is, I have to give 5 because it addresses the requirements of a narrative essay.
Organisation	The number of paragraphs is too many for such an event. But the transitions and the arrangement of the story is quite good. 4/5.
Coherence	It is fully coherent, flows smoothly. Explicit and implicit transitions are used. It is very professional. I do not trust the writer wrote himself, I am so so sure it is AI. 5/5.
Language Use (Mechanics)	There are no mistakes but it is too much for the task. The vocabulary is overcrowded. Choice of words is complex. 3/5.
Overall Feedback	I can give 17, I didn't give three points because it doesn't use a professional tone, rather, a poetic one. 

Essay Type	Narrative Essay 7
Title of the Essay	A Day of Unfortunate Events
Task Response	5- The essay effectively addresses the topic and the task. It also gives an important lesson and it has a silver lining at the end. So, it gives an appropriate response to the task.
Organisation	5- It is well organized and well developed . There are appropriate and sufficient explanations.
Coherence	4- The essay displays unity, progression, and coherence though some ideas could be connected better with more explanations.
Language Use (Mechanics)	5- There are not any noticeable mistakes in terms of grammar and the vocabulary used is diverse.
Overall Feedback	19-The essay is generally well organized, coherent and nice to read. 

Essay Type	Narrative Essay 7
Title of the Essay	A Day of Unfortunate Events
Task Response	The topic described in the title has been presented in a very coherent manner. 5/5
Organisation	The organization is quite adequate and orderly. The arrangement of the paragraphs is exactly as it should be, which adds fluidity to the narrative. 5/5
Coherence	The transitions between sentences are quite successful. Paragraph transitions have been used as scene changes and are interconnected. 5/5
Language Use (Mechanics)	There are some minor grammar mistakes; otherwise, the expressions used are varied. 4/5
Overall Feedback	The narration is excellent, particularly as it can be easily visualized in the reader's mind. The concluding section effectively ends the described situation with a profound life lesson. 19/20

1. Task Response

The essay effectively addresses the task by providing a detailed narrative of a challenging day. The writer recounts a series of unfortunate events and reflects on the experience, highlighting personal growth and resilience. The essay provides a clear beginning, middle, and end, with a reflective conclusion that ties the narrative together.

Score: 5/5

2. Organisation

The essay is well-organized, with a clear introduction, body, and conclusion. Events are presented in chronological order, making it easy for the reader to follow the narrative. Each paragraph focuses on a specific part of the day, enhancing readability. However, some transitions between paragraphs could be smoother to maintain the flow.

Score: 4/5

3. Coherence

The essay is coherent, with each event logically following the previous one. However, some transitions between sentences and paragraphs could be more explicit to improve the overall flow. For example, using transitional phrases like "After that," "Meanwhile," or "Later" would help guide the reader through the sequence of events more smoothly.

Score: 4/5

4. Language Use

The language use in the essay is generally clear and effective, but there are several grammatical errors and awkward phrases. Examples include "I sprained my foot" (should be "As I raced out the door, I sprained my foot") and "Despite this, I walked and took the medicine prescribed by the doctor" (could be rephrased for clarity). There are also minor punctuation issues. Improving grammar, sentence structure, and word choice would enhance the clarity and readability of the essay.

Score: 3/5

5. Overall Feedback

Overall, the essay provides an engaging and well-structured narrative of a challenging day. The task response is strong, and the organization is effective, with

a clear introduction, body, and conclusion. Coherence is generally good, but some transitions could be smoother. Language use needs improvement in terms of grammar, sentence structure, and word choice. For a first-grade ELT student, this essay demonstrates a good understanding of narrative structure and personal reflection.

Overall Score: 16/20

Detailed Suggestions for Improvement:

Task Response: Maintain the strong narrative and reflective elements. Consider adding more specific details about how each event affected you emotionally to deepen the reader's connection to the story.

Organisation: Ensure that each paragraph transitions smoothly to the next. Use more transitional phrases to guide the reader through the narrative.

Coherence: Improve the flow between sentences and paragraphs with clearer transitions. This will help the reader follow the sequence of events more easily.

Language Use: Focus on correcting grammatical errors and awkward phrases. Use grammar checking tools or seek feedback from a peer or instructor to identify and correct mistakes. Enhance sentence structure and word choice for better clarity and readability.

Overall Feedback: Continue practicing writing narratives and seek regular feedback to improve. Pay attention to grammar and coherence to make your essays more polished and engaging.

1. Task Response: 4/5

The essay effectively addresses the task of narrating a personal experience. The writer provides a detailed account of a day when everything went wrong, including the events and their consequences. The reflection on the lessons learned is present but could be expanded for a higher score.

2. Organisation: 4/5

The essay is well-organized with a clear structure. It begins with an introduction to the day, followed by a chronological recount of events, and concludes with reflections. Some transitions between paragraphs could be smoother to enhance the overall flow.

3. Coherence: 4/5

The essay maintains coherence by consistently focusing on the theme of a day when everything went wrong. The writer uses appropriate transitions and connectors to link ideas and events. However, there are a few areas where the narrative could be more tightly connected to improve clarity.

4. Language Use: 3/5

The language use is generally effective, with a good range of vocabulary and sentence structures. There are several grammatical errors and awkward phrasings that slightly impede understanding. The writer demonstrates a solid grasp of English, but there is room for improvement in accuracy and fluency.

5. Overall Feedback: 15/20

Overall, the essay is a strong narrative that effectively fulfills the task requirements. The writer's ability to organize and present their thoughts clearly, along with their reflective insights, makes for an engaging read. With some minor improvements in language use and coherence, the essay could achieve an even higher level of proficiency.

Great job! Keep practicing, and you'll continue to improve your writing skills. 😊

1. Task Response

Score: 4 out of 5 The essay effectively addresses the prompt of a day when everything went wrong. The writer details a series of unfortunate events, from spraining their foot to losing their phone, creating a relatable and engaging narrative.

2. Organization

Score: 3 out of 5 The essay is generally well-organized, following a chronological sequence. However, there could be more clarity in the transitions between events.

3. Coherence

Score: 3 out of 5 The essay maintains coherence, but there are a few instances where the flow of ideas could be improved. For example, the transition from the initial sprained foot to the lost phone could be smoother.

4. Language Use

Score: 3 out of 5 The writer's language is clear and understandable. However, there are some grammatical errors and awkward sentence constructions that could be corrected.

5. Overall Feedback

Overall Score: 13 out of 20 The essay is a personal narrative that effectively conveys a challenging day. The writer demonstrates a good understanding of the task and provides a relatable account of the events. However, there are areas for improvement in organization and language use.

Suggestions for Improvement:

Enhance Transitions: Use stronger transition words and phrases to improve the flow between events. For example, instead of simply stating "As if sensing my despair, the universe conspired to add more obstacles," you could use a transition like "Unfortunately, my troubles were far from over."

Explore Emotional Impact: Delve deeper into the emotional impact of the events on the writer. How did the writer feel when they sprained their foot? What were their anxieties about losing their phone?

Consider the Broader Implications: Explore how this experience may have affected the writer's perspective on life or their approach to handling challenges.

By incorporating these suggestions, the writer can enhance the overall quality of the essay and make it even more impactful.