Asian Journal of Distance Education

Comparative Analysis of Human Graders and AI in Assessing Secondary School EFL Journal Writing

Seval Kemal, Ayşegül Liman-Kaban

Abstract: This study conducts a comprehensive analysis of the assessment of journal writing in English as a Foreign Language (EFL) at the secondary school level, comparing the performance of a Generative Artificial Intelligence (GenAI) platform with two human graders. Employing a convergent parallel mixed methods design, quantitative data were collected from 389 assignments of 91 students in a private school in Istanbul during the first semester of the 2023-2024 academic year, evaluated by both the GenAI platform and human graders. Qualitative data involved analyzing feedback from both sources. The study aimed to compare grading performance, assess the GenAI platform's consistency and effectiveness, and examine feedback quality. Results indicated a high level of agreement between the GenAI platform and human graders, suggesting the GenAI platform can effectively simulate an English teacher's role in an EFL context. Limitations include the restricted sample size, the study's specific context, and potential variability in evaluations. Findings highlight the potential for integrating GenAI in EFL assessment, though human feedback remains crucial for personalized and emotionally supportive feedback. The conclusion emphasizes the GenAI platform's promise in enhancing feedback efficiency and comprehensiveness, while recommending future research to explore evaluation criteria, long-term impacts, and ethical considerations.

Keywords: AI-driven assessment, human vs. AI grading, generative AI, assessment in education, EFL assessment, feedback quality, AI in education, writing assessment, formative feedback, automated scoring.

Highlights

What is already known about this topic:

- Human feedback is crucial for personalized and emotionally supportive learning.
- AI platforms can provide consistent and detailed feedback in educational settings.

What this paper contributes:

- Demonstrates the effectiveness of a GenAI platform in assessing EFL journal writing.
- Shows high agreement between GenAI platform and human graders in EFL assessments.
- Highlights the potential for integrating AI in EFL writing assessments.

Implications for theory, practice and/or policy:

- Encourages the use of GenAI platforms to complement human graders in EFL contexts.
- Suggests refining evaluation criteria for EFL assessments using AI.
- Recommends future research on long-term impacts and ethical considerations of AI in education.

Published by Asian Society for Open and Distance Education (ASODE), Japan
ISSN 1347-9008 http://www.asianjde.com/

Introduction

Digital technologies have dramatically transformed the educational landscape, as highlighted by Timotheou et al. (2023). One of the most significant advancements is the integration of artificial intelligence (AI). As described by Cao (2023), AI encompasses the science of designing intelligent machines that emulate human behaviors, aiming to replicate and enhance human capabilities (Naqvi, 2020). AI's impact, particularly in education, has reshaped learning environments in profound ways (Bozkurt, 2023).

Among the various AI applications, GenAI has emerged as a crucial tool in language education. GenAI, a subset of AI, autonomously generates content in response to specific prompts (Miao & Holmes, 2023). Large language models (LLMs), such as OpenAI's GPT-3, Google's Bard (later Gemini), and Microsoft's Bing, have revolutionized how education, particularly language learning, is approached. These models demonstrate capabilities like addressing queries, challenging flawed assumptions, acknowledging mistakes, and producing high-quality written content, often exceeding the skills of many human counterparts (Elkins & Chun, 2020).

In the context of English as a Foreign Language (EFL) education, AI has been applied in several domains, including automated written corrective feedback (Koltovskaia, 2020). However, research on the efficacy of LLMs in assessing writing, particularly at the K-12 level, remains scarce (Agostini, 2024; Escalante et al., 2023; Algaraady & Mahyoob, 2023). This study seeks to address this gap by comparing the performance of a GenAI platform against two human graders in evaluating secondary school students' EFL journal writing.

This research adopts a convergent parallel mixed methods design, integrating both quantitative and qualitative analyses to assess the quality of feedback provided by GenAI and human graders. The study further examines how varying prompts influence the interaction between students and GenAI and investigates GenAI's potential to simulate the role of an English teacher in a classroom setting.

The findings of this study will contribute to the discourse on integrating AI into education, particularly in assessing writing skills. By evaluating GenAI's potential alongside human graders, the study aims to offer insights into the effectiveness of AI in educational assessment and inform educators, policymakers, and researchers on the strengths and limitations of AI-enhanced learning environments.

This research aims to answer the following questions:

1. How do the quality and characteristics of feedback from ChatGPT and human evaluators differ concerning content, language, and organization in student essays?
   1.1. Is there a statistically significant difference between ChatGPT and human graders' evaluations?
   1.2. Is there a statistically significant correlation between ChatGPT and human graders' evaluations?
2. How does the performance of the GenAI platform compare to human graders in the assessment of EFL journal writing in secondary school?
   2.1. How do human graders and GenAI provide feedback in terms of praise or a supportive tone?
   2.2. How do human graders and GenAI provide corrective feedback for mistakes?
   2.3. How do human graders and GenAI provide guidance on student papers?
   2.4. How do human graders and GenAI encourage students to improve in their next papers?

Literature

Technology and AI in Language Learning

The integration of technology into education, particularly in language learning, has transformed traditional approaches. Technology enables personalized, adaptive learning experiences, allowing students to receive real-time feedback and engage in interactive tasks that align with their learning needs (Holmes et al., 2019). Among these technologies, Artificial Intelligence (AI) has emerged as a significant tool, offering innovative solutions to complex educational challenges.

The concept of AI was first introduced during the 1956 Dartmouth Summer Research Project, led by John McCarthy and colleagues. AI's ability to mimic human intelligence has evolved significantly since its inception. Alan Turing, one of the pioneering figures in computing, posited that machines could eventually think in ways akin to humans (Turing, 1950). Today, AI spans across various fields, but its role in education has gained particular attention for its potential to personalize and enhance learning.

In language learning, AI facilitates the automation of tasks like grading, content generation, and formative assessments. By analyzing patterns in student responses, AI can provide timely feedback and tailor instructional materials, ultimately promoting a more individualized learning experience. The use of AI in English as a Foreign Language (EFL) settings, for example, has revolutionized the way teachers assess student writing and offer feedback.

Artificial Intelligence in Education and Assessment

AI's application in education has transformed how assessment and feedback are delivered, particularly in language learning. AI technologies such as automated grading systems, natural language processing (NLP), and machine learning (ML) algorithms are increasingly being employed to assess students' writing skills (Roll & Wylie, 2016). These tools analyze student texts, evaluate language use, and provide feedback, simulating a human grader’s evaluation process.

Generative AI (GenAI), such as the Generative Pre-trained Transformer (GPT) models, offers additional benefits by generating feedback based on predefined rubrics or prompts. For example, models like GPT-4 process large datasets of language inputs to deliver feedback that is contextual, relevant, and often as accurate as human evaluation. Such AI systems improve efficiency, allowing teachers to focus more on instruction and less on manual grading.

In addition, AI enhances assessment reliability by standardizing the evaluation process. Human graders may inadvertently introduce bias into assessments, while AI can consistently apply predefined criteria to student writing, reducing subjectivity (Baidoo-anu & Owusu Ansah, 2023). This demonstrates AI's potential to replicate or complement traditional grading methods.

Comparison of Human and AI Grading

While AI holds significant promise in education, a critical area of study is the comparison between human and AI grading, particularly in the context of language assessment. Human graders bring nuanced understanding and empathy to evaluations, considering elements like creativity and context. AI, on the other hand, excels in processing large amounts of data efficiently, providing consistent feedback based on predefined rubrics.

The results indicate a notable alignment between the grades provided by the teacher and those generated by ChatGPT for the evaluated writing essays, suggesting that AI tools like ChatGPT can reliably assess language proficiency and writing quality in educational contexts (Roll & Wylie, 2016)**.** However, differences emerge in more subjective areas such as tone, argumentation, and creativity. For instance, AI may struggle with understanding cultural or contextual nuances in writing, something human graders are naturally attuned to (Chassignol et al., 2018). This disparity highlights the importance of human oversight in AI-augmented grading systems.

The thematic analysis of feedback from AI and human graders in this study reveals that AI provides more corrective feedback, focusing on grammar and structure, while human graders offer more praise and encouragement. This aligns with findings from Baidoo-anu and Owusu Ansah (2023), who noted that AI systems are more focused on technical accuracy, while human evaluators provide feedback that fosters deeper student engagement and motivation. Thus, a blended approach combining AI efficiency with human empathy may offer the best of both worlds.

Relevance to Secondary Education and EFL

In secondary education, particularly in EFL settings, AI has shown great potential to augment learning and assessment practices. Students in EFL classrooms benefit from AI's ability to provide immediate, detailed feedback on writing assignments, enabling them to address errors more quickly and refine their language skills over time. In this study, involving 91 students in 5th and 6th grade, AI feedback on journal writing was compared to feedback provided by human graders. This comparison sheds light on how AI can support language acquisition at an early educational stage by complementing traditional teacher evaluations.

AI can enhance writing instruction in EFL by offering personalized feedback that caters to individual student needs. For example, in formative assessments, AI tools can identify patterns in language use, help students understand recurring errors, and provide guidance on improving linguistic accuracy (UNESCO, 2023). The potential to integrate AI into EFL classrooms promises a more adaptive and responsive learning environment for students, fostering more effective language acquisition.

However, the integration of AI into secondary education also presents challenges. Teachers must understand the limitations of AI systems, such as their inability to grasp deeper cultural or emotional aspects of language. While AI can efficiently grade essays and identify grammatical mistakes, it cannot yet replicate the nuanced feedback provided by human graders, which is crucial in helping students develop critical thinking and creative skills (Roll & Wylie, 2016).

In conclusion, the integration of AI in education, particularly in language learning and assessment, offers numerous benefits. AI technologies streamline the assessment process, provide consistent feedback, and help teachers focus more on instructional tasks. However, AI systems have limitations, particularly in areas requiring deeper human insight. This study's comparison between AI and human feedback in secondary EFL education highlights the complementary roles these systems can play, emphasizing the need for a balanced approach that leverages the strengths of both AI and human evaluators.

Theoretical Background

In recent years, the integration of technology in education has transformed traditional teaching and learning methods. Artificial intelligence (AI) applications, in particular, have emerged as powerful tools with the potential to revolutionize the field of education. AI technologies, such as LLMs and GenAI, are

increasingly being used to support language learning and teaching, including the assessment of writing skills.

Writer(s)-Within-Community Model of Writing (WWC)

The theoretical framework of this study draws on the Writer(s)-Within-Community Model of Writing (WWC). As posited by Graham (2018), the foundation of the WWC model lies in the concept that writing is a communal endeavor, embedded within particular settings known as writing communities. According to Graham (2018), within the WWC model, a writing community is described as a cohort of individuals who possess shared goals and assumptions, utilizing writing as a tool to accomplish their objectives. The objectives and assumptions of these communities can exhibit significant diversity, ranging from clearly stated to implicitly understood, and may evolve over time, exhibiting a range of states from stable to emergent or shifting. Writing collaboratives consist of writers, collaborators, and readers who function as an audience (Cameron, Hunt, & Linton, 1996). In a writing community located in a school, this may be a teacher or a peer who serves as a supervisor (Graham, 2018). A writing community may have a top-down organizational model, as is often the case in schools, where an authority figure takes on the role of teacher (Graham, 2018). The task of writing involves five production processes: conceptualization, ideation, translation, transcription, and reconceptualization. Conceptualization includes generating a conceptual representation of the task; ideation includes generating content from memory or external sources; translation involves transforming content into sentences that convey intended meanings; transcribing involves writing over printed or digital text; and reconceptualization involves revision (Steiss et al., 2024). The current study concentrated on reconceptualization since it is possible to enhance students' ability to reconceptualize the texts they write by giving them high-quality formative feedback. This kind of feedback helps students improve their writing.

Importance of Feedback in Writing Instruction

Warschauer and Ware (2006) highlight that effective writing instruction often involves teachers providing frequent individual feedback on multiple drafts for each student. This iterative process helps students refine their writing skills through targeted suggestions and corrections. However, they emphasize that such personalized feedback is time-intensive, posing challenges in environments with large student-to-teacher ratios.

Research underscores the critical role of timely and specific feedback in fostering writing development. According to Black and William (2009), infrequent feedback can hinder the learning process by reducing opportunities for students to address their errors and improve. Immediate responses, tailored to students' needs, are vital for their ability to internalize and apply feedback effectively.

In addition to timeliness, the nature of feedback significantly influences its effectiveness. Hattie and Timperley (2007) propose a model emphasizing feedback's role in clarifying goals, providing actionable suggestions, and motivating students to bridge gaps in their performance. Their research concludes that formative feedback enhances writing quality by enabling students to focus on the aspects of their writing that need improvement.

Automated writing evaluation (AWE) tools, such as ChatGPT, have emerged as potential solutions to the time constraints of traditional feedback. However, the balance between automated and human feedback remains a critical area of inquiry, as students often benefit from the nuanced and context-specific insights provided by human graders (Steiss et al., 2024).

Moreover, feedback's role extends beyond addressing linguistic errors; it shapes students' attitudes toward writing and learning. Deci and Ryan's (2000) Self-Determination Theory indicates that supportive feedback fosters intrinsic motivation, leading to greater persistence and engagement in writing tasks.

In conclusion, investigating innovative feedback mechanisms, such as integrating AWE tools like ChatGPT, could enhance the efficiency and effectiveness of writing instruction. Combining automated and human feedback approaches may offer a comprehensive solution that leverages the strengths of both, ultimately improving learning outcomes.

Context of the Study

In the context of this study, the writing community consists of 91 English as a Foreign Language middle school students from 5th and 6th grades in Istanbul. Their shared goal is to enhance their writing skills, achieved through the bi-weekly assignment of journal topics. Teachers provide formative feedback on these assignments, focusing on three key criteria: content, language, and organization. This feedback is graded on a scale of 1 to 5, with 1 being the lowest and 5 being the highest. Additionally, to assess the consistency of feedback within this framework, ChatGPT 3.5 (From version of October, 2023 to June, 2024) is utilized alongside two human graders. This approach aligns with the WWC model, as it recognizes writing as a collaborative and social process involving multiple participants, including students, teachers, and technology (ChatGPT 3.5). By integrating this model into the study, we aim to explore the effectiveness of utilizing ChatGPT 3.5 alongside human graders in providing consistent and constructive feedback within a writing community, ultimately enhancing the students' writing capabilities.

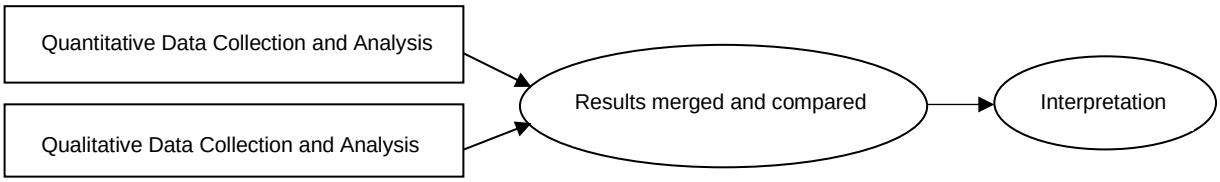
Methodology

The researcher is guided in organizing and carrying out the study in a way that is most likely to accomplish the desired outcome by methodology and research design. It serves as a guide for carrying out the research (Burns & Grove, 2005). This part describes the methods used in this study, with particular attention on data analysis, data collection tools, and population and sample selection. The research employs a convergent parallel mixed methods approach, wherein quantitative and qualitative data are concurrently collected and analyzed (Creswell & Plano Clark, 2018). This design allows for a comprehensive assessment of the effectiveness and consistency of AI in grading student papers in comparison to human graders.

Research Model/Design

In order to compare or integrate the results, the researcher uses a convergent parallel mixed methods methodology, in which two separate datasets—one qualitative and one quantitative—are collected and analyzed separately before being combined (Creswell & Plano Clark, 2018). This design, aimed at gaining varied yet complementary data on the same subject (Morse, 1991), seeks to amalgamate the strengths and weaknesses of quantitative and qualitative methodologies (Patton, 1990), such as the depth of detail and subjective interpretation in qualitative research, and the larger sample size and objective measures in quantitative research. Researchers employ this approach when seeking a comprehensive understanding of the research problem, juxtaposing quantitative statistical data with qualitative insights (Creswell & Plano Clark, 2018). However, the simultaneous collection, analysis, and merging of quantitative and qualitative data in the convergent design can pose philosophical challenges, leading researchers to favor pragmatism as an overarching worldview over attempts to reconcile different paradigms (Creswell & Plano Clark, 2018).

Figure 1. Convergent Parallel Mixed Method Methodology



In the quantitative component**,** numerical data was collected to quantify the effectiveness of ChatGPT 3.5 in grading EFL writing. This data included scores assigned by both human graders and the ChatGPT 3.5 evaluated under *content, language and organization* criteria as well as statistical analyses comparing the two sets of scores.

In the qualitative component, textual data was collected to understand the quality and characteristics of feedback provided by both human graders and the AI system. This data included examples of feedback provided for student journals, as well as thematic analyses coded as praise, correc-tion, guidance and encouragement to identify common patterns and themes in the feedback.

The quantitative and qualitative data were collected and analyzed independently. However, during the integration phase, the findings from both components were compared to identify convergent or divergent patterns. By offering a more thorough and sophisticated knowledge of the research issues, this data integration improved the study's validity and dependability (Teddlie & Tashakkori, 2009). The results from both components suggest a convergence of evaluations between human graders and the GenAI platform in terms of language, content, and organization, also in the nature and extent of feedback provided.

Data Collecting Tools

In this study, several data collection tools were employed to ensure comprehensive and reliable data for both quantitative and qualitative analyses. The primary source of data comprised student essays, with 91 middle school students from 5th and 6th grades in Istanbul submitting bi-weekly journal assignments. These essays were assessed by two experienced EFL teachers and ChatGPT 3.5. The human graders provided detailed feedback on the essays, focusing on content, language, and organization, using a standardized evaluation rubric. Given that Cambridge University Press books are used at the school and are internationally recognized, the rubric for A2 Key for Schools was employed to maintain consistency and reliability in evaluations. This rubric, featuring descriptors for each criterion and a scoring scale from 1 to 5, was also used by ChatGPT 3.5 to ensure uniformity across assessments.

Table 1. Writing Assessment subscales for A2 Key for Schools

A2	Content	Organisation	Language
5	All content is relevant to the task. Target reader is fully informed.	Text is connected and coherent, using basic linking words and a limited number of cohesive devices.	Uses everyday vocabulary generally appropriately, while occasionally overusing certain lexis. Uses simple grammatical forms with a good degree of control. While errors are noticeable, meaning can still be determined.
4	Performance shares features of Band 3 and 5.		
3	Minor irrelevances and/or omissions may be present. Target reader is, on the whole, informed.	Text is connected using basic, high-frequency linking words.	Uses basic vocabulary reasonably appropriately. Uses simple grammatical forms with some degree of control. Errors may impede meaning at times.
2	Performance shares features of Band 1 and 3.		
1	Irrelevances and misinterpretation of the task may be present. Target reader is minimally informed.	Production unlikely to be connected, though punctuation and simple connectors (i.e. ‘and’) may on occasion be used.	Produces basic vocabulary of isolated words and phrases. Produces few simple grammatical forms with only limited control.

Note. Adapted from "Teacher Guide for Writing A2 Key for Schools," by Cambridge English, 2020.

To facilitate quantitative analysis, statistical software such as SPSS was utilized to analyze the evaluation data, determining significant differences and correlations between human and AI feedback. These data collection tools collectively provided a robust framework for comparing the quality, consistency, and constructiveness of feedback from human graders and the GenAI, addressing the research questions of this study.

Sampling or Study Group

The study used a purposive sampling method, selecting 91 middle school students from 5th and 6th grades in Istanbul, Turkey selected to represent a diverse and typical group of English as a Foreign Language (EFL) learners at the middle school level. This approach ensured the participants were specifically chosen based on predetermined criteria, such as their age, grade level, and perhaps their access to English language learning programs. Unlike random sampling, purposive sampling allows researchers to focus on a group that is expected to provide the most relevant and insightful data for the research objectives. In this case, the target population of middle school students was chosen because they are at a critical stage in developing foundational writing and language skills, making them an ideal group for studying the effects of AI-assisted feedback in educational settings.

Table 2. Demographic Information of Participants

Grade	Gender	Number of Participants	N %	Age
5th	Girls	18	19.78	10
5th	Boys	34	37.36	10
6th	Girls	17	18.68	11
6th	Boys	22	24.18	10
Total		91	100	

The participants were part of an ongoing EFL program that incorporated bi-weekly journal writing assignments. These assignments provided a substantial dataset for analyzing and comparing feedback from human graders and ChatGPT 3.5. The human graders in this study were two experienced EFL teachers with extensive backgrounds in teaching and assessing English writing skills.

Table 3. Demographic and Professional Background of Assessors

Assessor	Age	Experience	Teaching Experience	Certificate
Human Grader 1	43	20	Levels 1-8 at K-12 schools	No
Human Grader 2	40	20	K-5 to adult learners	Yes
GenAI	*	*	*	*

The selection of this specific sample and participant group was intentional to ensure that the study could provide meaningful insights into the effectiveness of AI-assisted feedback in a real-world educational context. By focusing on middle school students actively engaged in EFL learning and using a recognized assessment rubric, the study aimed to rigorously examine the research questions and contribute valuable findings to the field of educational technology and language assessment.

Research Procedures

First, writing topics and the rubric were determined at a departmental meeting. In early October 2023, all classes were provided with writing workshops and journal booklets by their respective teachers. Following this, a pilot class was selected to represent the study. The initial journals from this class were evaluated by two human graders and three different Generative Artificial Intelligence platforms: Google's Bard which is Gemini as of July 2024, Meta's Llama, and OpenAI's ChatGPT 3.5 version of October 2023. The prompts were varied three times to identify the most effective prompt. Subsequently, other classes were included in the project. The researcher began transcribing journals into a Word document, maintaining the original formatting, including any spelling mistakes or capitalization used by the students. These transcriptions were then sent to human graders via email, with student names anonymized and codes assigned to each student to protect their identities. After each submission, all papers were assessed, scored, and provided feedback by teachers. Simultaneously, the feedback was saved to be coded under specific themes. The journal writing project commenced in October 2023 and concluded in January 2024.

Piloting

Following consultations with a subject matter expert, an experimental procedure involving Google Bard and ChatGPT 3.5 was commenced in October 2023. Initially, a sample classroom was chosen as a test classroom, where the preliminary assessments of the journal were executed using Google Bard. After employing Bard, the assessment of documents was conducted through the utilization of ChatGPT 3.5. Following the examination of five documents, an inquiry was made to ChatGPT 3.5 to ascertain whether it still possessed recollection of the subject matter and the evaluation criteria, given the existence of some speculation concerning its memory. Though a rating scale ranging from 0 to 5 was requested, a rating system spanning from 1 to 6 was generated. It then reassessed, as instructed and using the appropriate assessment criteria, the five individuals it had previously assessed. However, the results revealed inconsistent findings. In particular, it assessed the third and fifth students, whom it had previously classified as scoring two, now as four. At this point, the marking became misleading.

Next, a sample paper written by a teacher that met the criteria for receiving the highest possible score was presented. ChatGPT 3.5 was tasked with evaluating this piece of work, and it assigned a score of 4. When asked about the errors or missing elements, ChatGPT 3.5 provided feedback that did not align with the actual evaluation criteria. It was then asked to produce a text worthy of 5 marks, which it generated at a C1 level, far above the students' actual proficiency. After clarifying the students' age, competence, and educational background, ChatGPT 3.5 produced a more appropriately leveled text. However, it inconsistently rated the teacher's sample text, initially rated lower than deserved, as a 5 when directly compared to its own text. These inconsistencies led to further skepticism about ChatGPT 3.5’s ability to accurately rate perfect papers.

Following this, Bard was also tested using the teacher's sample text. It initially scored the sample as 4, like ChatGPT 3.5, and produced a sample text at the B1 level when prompted. After providing additional context about the students' level, Bard eventually rated the sample text correctly. The performance of Meta's Llama was also briefly assessed, but due to its tendency to create its own rubric, partial evaluations, and memory deletion issues, it was quickly excluded from further consideration.

The preliminary pilot investigation highlighted the importance of prompt design in utilizing generative artificial intelligence effectively. Based on initial testing, prompts were reorganized to exclude zero scores, incorporate learning objectives, and present exemplary samples. ChatGPT 3.5 was consulted on the best practices for creating prompts, emphasizing that effective prompting involves explicitness, avoiding multiple questions simultaneously, and using complete sentences. As a result of these insights, a conversational approach was deemed more effective, supported by additional scholarly research. Bozkurt and Sharma (2023) underscore that strategically approaching conversational generative AI with a well-defined purpose, tone, role, and context allows for the establishment of prompt-based conversational pedagogy. This framework facilitates meaningful communication and interaction, enhancing the teaching and learning experience. In alignment with this principle, the final pilot study in this research explored conversational cascading prompts, leading to the selection of ChatGPT 3.5 as the most consistent and reliable GenAI platform for further research.

Data Analysis

The data analysis procedures for this study involved both quantitative and qualitative approaches. For quantitative analysis, the Statistical Package for the Social Sciences (SPSS) software was used. First, the quantitative data collected from the human graders and ChatGPT 3.5 were entered into SPSS for statistical analysis. Descriptive statistics summarized the data, including means, standard deviations, and frequencies. Next, inferential statistics, such as t-tests and correlations, were used to compare the evaluations between human graders and ChatGPT 3.5. This analysis aimed to determine if there were significant differences or correlations in the evaluations.

For qualitative analysis, thematic analysis was employed. Feedback provided by human graders and ChatGPT 3.5 was collected in an Excel document, including students' work. Themes were created to categorize the feedback into praise, correction, guidance, and encouragement. A checklist was developed for ChatGPT 3.5 and each grader to determine the presence and frequency of these themes. Additionally, word counts were recorded as graders were not restricted to a specific word limit.

The analysis process began with a preliminary reading of the feedback. Open coding was performed in the second stage, with different codes assigned colors for easy recognition. Open coding involves breaking down data into distinct parts and closely examining them to identify and categorize key concepts and themes (Strauss & Corbin, 1998). The coding was then reviewed to ensure alignment with established themes. This qualitative data analysis focused on feedback from one class, identifying patterns and trends in feedback provided by human graders and ChatGPT 3.5.

Overall, the combination of quantitative and qualitative data analysis provided a comprehensive understanding of the feedback dynamics between human graders and ChatGPT 3.5. The quantitative analysis, conducted using SPSS, identified significant differences and correlations in the evaluations, providing a statistical foundation for the study's findings. Concurrently, the qualitative analysis, through meticulous coding and thematic categorization, offered nuanced insights into the nature and quality of the feedback. Integrating both approaches allowed for a holistic assessment of feedback mechanisms, ultimately contributing to a deeper understanding of AI's potential roles in educational contexts.

Validity and Reliability of the Evaluation Criteria

Ensuring the validity and reliability of the evaluation criteria is a critical aspect of this study to establish the credibility of the findings. As defined by Baykal (2015), validity refers to the degree to which the assessment measures what it is intended to measure and maintains relevance and conformity. Reliability, on the other hand, pertains to the consistency and stability of the measurements over time and across evaluators (Jacobs, & Sorensen, 2010).

Validity Assessment

To confirm the validity of the evaluation criteria (organization, language, and content), the rubric used in this study was adapted from the Cambridge Assessment A2 Key for Schools, a well-established framework in language assessment.

- *Content Validity:* The rubric was reviewed and refined by five experts in the field of EFL and educational assessment to ensure that each criterion accurately represented the skills being measured (organization, language, and content). Adjustments were made based on their feedback to better align with the writing tasks and the proficiency level of the students.
- *Construct Validity:* The evaluation criteria were tested during a pilot study involving a subset of student writings. This pilot analysis confirmed that the rubric effectively captured the intended aspects of writing quality.

Reliability Assessment

Reliability was assessed to ensure consistent application of the evaluation criteria across different raters and over time:

- Interrater Reliability: Human graders were trained on the rubric, and their scoring was compared using correlation analysis to assess the consistency of their evaluations. The correlation coefficients were calculated using SPSS to determine the level of agreement among the graders.
- Internal Consistency: Cronbach's alpha was calculated for the scores given to each criterion (organization, language, and content) to assess the internal consistency of the rubric and the overall reliability of the grading process. A Cronbach's alpha score of 0.7 or higher was considered acceptable.

Limitations

During the first semester of the 2023-2024 academic year, data was collected from 91 students in a private school in Istanbul, which is a slightly restricted sample size that may limit the generalizability of the results. Although initially intended to involve 110 students, absences and non-submissions reduced the final sample size. This study's findings are specific to K-12 EFL journal writing and may not apply to other educational settings, populations, or tasks. Conducted over a 13-week period, the study faced time constraints and the cancellation of the last journal due to exam weeks, potentially affecting the results. Human grading variability and subjective factors, such as tiredness and mood, also posed limitations, despite efforts to ensure consistency. Additionally, while using ChatGPT 3.5 for feedback, there were instances where the AI did not strictly adhere to the rubric and required reminders of evaluation criteria. The study's reliance on a specific platform means the findings may not be applicable to other technologies. Moreover, biases based on overall student performance and the ability to identify students despite blinded coding could impact objectivity. Lastly, the specific prompts used might not represent the full range of possible prompts, and other research gaps not addressed in this study could further limit its conclusions. Therefore, while providing valuable insights, these findings should be interpreted cautiously, and future research should aim to address these limitations.

Results

Quantitative Data Analysis

The study found high concurrent validity between ChatGPT 3.5 and human graders' assessments of student writing, indicating consistent language, content, and organization. However, construct validity

revealed high redundancy in variables like GPTLNG, GPTCON, and GPTORG, suggesting they may not offer unique information. Despite this, high interjudge reliability was observed between human graders and the ChatGPT 3.5 indicating consistent and reliable evaluations. This enhances the credibility and accuracy of the grading process, despite potential redundancies in the evaluation criteria.

In the context of the study, the variables JDGORG, JDGLNG, JDGCON, GPTORG, GPTLNG, and GPTCON were used to represent specific aspects of the evaluation criteria for student writing, assessed by both human graders (judges) and AI (ChatGPT 3.5). These variables were developed to facilitate a quantitative comparison of the performance of human graders and AI in assessing student work. Here is a consolidated explanation of these variables:

JDGORG (Judge: Organization): This variable represents the human graders' evaluation of the organizational structure of the students' writing. Human graders analyzed how well the text was structured, focusing on the logical flow of ideas, coherence, proper paragraphing, and transitions. Scores were assigned based on the clarity and effectiveness of the introduction, body, and conclusion.

JDGLNG (Judge: Language): This variable refers to the assessment of language use by the human graders. It included an analysis of grammar, vocabulary, sentence structure, and overall linguistic accuracy. Scores were reflective of the student's command of English, considering aspects like grammatical correctness, vocabulary appropriateness, and fluency.

JDGCON (Judge: Content): This variable captures the human graders' evaluation of the content in the writing. Graders assessed the relevance, depth, originality, and overall insightfulness of the ideas presented in response to the given prompts. The focus was on how effectively students addressed the topic and developed their arguments.

GPTORG (ChatGPT 3.5: Organization): This variable corresponds to ChatGPT's evaluation of the organizational structure of the students' work. Similar to human grading, the AI assessed aspects like logical flow, paragraph coherence, and structural clarity, assigning scores based on these criteria.

GPTLNG (ChatGPT 3.5: Language): This variable reflects the AI's assessment of linguistic quality, including grammar, vocabulary, and sentence construction. ChatGPT evaluated fluency, accuracy, and appropriateness of language, offering scores aligned with these factors.

GPTCON (ChatGPT 3.5: Content): This variable pertains to ChatGPT's assessment of content quality. The AI evaluated the relevance and depth of ideas, originality, and the alignment of the response with the given prompt. Scores were based on how well students articulated and supported their ideas.

Assessment and Data Collection Process: Both human graders and ChatGPT 3.5 were guided by rubrics detailing criteria for organization, language, and content. Human graders underwent training to ensure consistent and reliable evaluations. Each student's work was independently assessed, and the scores for JDGORG, JDGLNG, and JDGCON were recorded. For AI assessment, ChatGPT provided scores for GPTORG, GPTLNG, and GPTCON based on the same rubrics. The collected data were subsequently analyzed using SPSS, allowing for a quantitative comparison of human and AI evaluations across these variables.

Table 4. Kendall's Tau Correlation Analysis of Evaluation Criteria in EFL Journal Writing Assessment

Variable	Variable	Kendall's Tau	p	-log(p)
GPT-LANG	GPTCON	,874	,000	30,5
GPTORG	GPTCON	,903	,000	32,6
GPTORG	GPTLNG	,917	,000	33,5
JDGLNG	JDGCON	,905	,000	34,4
JDGORG	JDGCON	,921	,000	35,7
JDGORG	JDGLNG	,927	,000	36,2
JDG1	GPT	,698	,000	21,2
JDG2	GPT	,699	,000	21,3
JDG2	JDG1	,933	,000	37,3

GPTCON: Generative Pre-trained Transformer Content; GPTLNG: Generative Pre-Trained Transformer Language; GPTORG: Generative Pre-trained Transformer Organization; JDGCON Judge Content; JDGLNG: Judge Language; JDGORG: Judge Organization

The study analyzed the correlation between evaluation criteria in English as a Foreign Language (EFL) journal writing assessment, comparing human graders (JDGORG, JDGLNG, JDGCON) and ChatGPT 3.5 (GPTORG, GPTLNG, GPTCON). The results showed strong positive correlations between the evaluation criteria for both groups, with organization being the most consistently rated aspect. However, moderate positive correlations were found between ChatGPT 3.5 and human graders, indicating moderate agreement but also differences in evaluation methods. The study concluded that both human graders and ChatGPT 3.5 provide consistent evaluations of organization, language, and content in EFL journal writing assessment.

Table 5. Kendall's Tau Correlation Coefficients and Sum Scores for Evaluation Criteria in EFL Journal Writing Assessment

Variable	Kendall's Tau	p	-log(p)	SUM
GRADER-ORGANIZATION	0,921	0,000	35,7	71,9
GRADER-LANGUAGE	0,905	0,000	34,4	70,7
GRADER-CONTENT	0,905	0,000	34,4	70,1
GPT-ORGANIZATION	0,903	0,000	32,6	66,1
GPT-LANGUAGE	0,874	0,000	30,5	64,1
GPT-CONTENT	0,874	0,000	30,5	63,1
GRADER2	0,699	0,000	21,3	58,6
GRADER1	0,698	0,000	21,2	58,6
GPT	0,698	0,000	21,2	42,5

GPTCON: Generative Pre-trained Transformer Content; GPTLNG: Generative Pre-Trained Transformer Language; GPTORG: Generative Pre-trained Transformer Organization; JDGCON Judge Content; JDGLNG: Judge Language; JDGORG: Judge Organization

The analysis revealed that students who completed all five journals had significantly higher mean scores in the GPT (46.57), JDG1 (60.24), and JDG2 (59.39) categories compared to those who did not complete all journals. This suggests a strong positive correlation between consistent engagement with journal assignments and academic performance. Several factors could contribute to this observed trend. Firstly, regular practice and completion of writing tasks likely enhance students' writing skills,

thereby leading to improved performance. The process of continuous writing helps in reinforcing language structures, vocabulary, and critical thinking skills, which are essential components of proficient writing. Moreover, frequent writing assignments provide more opportunities for feedback, enabling students to identify and correct their errors, and thus improve over time. The higher scores in the GPT category highlight the potential of ChatGPT 3.5 in supporting student learning. The consistent use of GPT for feedback might have provided students with timely and detailed corrections, which are crucial for learning and improvement (Black & William, 2009). The significant scores in JDG1 and JDG2 also underscore the importance of traditional human feedback, suggesting that the combination of ChatGPT 3.5 and human input might offer a comprehensive support system for students. However, these results also raise questions about accessibility and equity. Not all students may have the same level of access to resources or support needed to complete all journal assignments. Future research should investigate whether certain groups of students are disadvantaged by this requirement and explore strategies to provide additional support to those who struggle to complete their assignments.

In conclusion, the findings indicate that consistent completion of journal assignments is associated with higher academic performance. This emphasizes the need for educators to encourage regular engagement with writing tasks and consider integrating AI tools like ChatGPT 3.5 to provide timely and effective feedback. Furthermore, addressing potential inequities in assignment completion should be a priority to ensure all students have the opportunity to benefit from these learning activities.

Table 6. Descriptive Statistics of Grades Based on Completed Journals for GPT, GRADER1, and GRADER2

DELIVERY		N	Mean	St.Dev.	SE Mean
GPT	1,00	37	25,22	9,304	1,530
	2,00	54	46,57	8,336	1,134
GRADER1	1,00	37	31,08	14,297	2,350
	2,00	54	60,24	11,621	1,581
GRADER2	1,00	37	30,41	14,450	2,376
	2,00	54	59,39	12,134	1,651

The study found gender differences in grading outcomes across three grading methods: GPT, JDG1, and JDG2. Female participants had a higher mean rank in the GPT method (52.04) compared to male participants (42.22), indicating higher grades on average. In the JDG1 method (51.09), female participants received higher grades on average (42.82), and in the JDG2 method (50.36), female participants received higher grades on average (43.28). These results suggest that female participants tend to receive higher grades compared to male participants in various grading methods.

Table 7. Ranks of Participants Based on Completed Journals for GPT, JDG1, and JDG2 Grading Methods

Ranks				
DLVRY		N	Mean Rank	Sum of Ranks
GPT	1,00	37	21,43	793,00
	2,00	54	62,83	3393,00
	Total	91		
JDG1	1,00	37	22,50	832,50
	2,00	54	62,10	3353,50
	Total	91		
JDG2	1,00	37	23,03	852,00
	2,00	54	61,74	3334,00
	Total	91		

The analysis of variance (ANOVA) results indicate statistically significant differences in the mean scores assigned by the three grading methods (GPT, JDG1, JDG2) for evaluating student papers in English as a Foreign Language (EFL) journal writing assessment. These findings suggest that the choice of grading method significantly influences the scores students receive, highlighting the importance of selecting an appropriate evaluation approach in EFL education.

Table 8. ANOVA Results for Mean Scores of GPT, JDG1, and JDG2 Grading Methods in EFL Journal Writing Assessment

ANOVA						
		SS	df	MS	F	Sig.
GPT	Between Groups	1778,513	2	889,257	5,204	,007
JDG1	Between Groups	3165,351	2	1582,676	4,639	,012
JDG2	Between Groups	3611,343	2	1805,672	5,270	,007

Qualitative Data Analysis

In educational settings, feedback is essential in guiding student learning and enhancing their academic outcomes. Traditionally, feedback has been given by human instructors, but with advances in AI, platforms like ChatGPT 3.5 now provide an alternative. This qualitative analysis compares the feedback from ChatGPT 3.5 and two human graders (HG1 and HG2) on student essays, focusing on four main themes: praise, correction, guidance, and encouragement. The analysis also looks at the language

used, the organization of the feedback, and its completeness. This comparison aims to highlight the strengths and limitations of AI-generated feedback.

Table 9. Frequency and Percentage Distribution of Feedback Themes by GPT and Human Graders

Themes	GPT Frequency	GPT Percentage	HG1 Frequency	HG1 Percentage	HG2 Frequency	HG2 Percentage
Praise	33	23.08%	28	31.46%	38	45.24%
Correction	22	15.38%	12	13.48%	14	16.67%
Guidance	42	29.37%	24	26.97%	15	17.86%
Encouragement	46	32.17%	25	28.09%	17	20.24%

In comparing praise, ChatGPT 3.5 feedback is more structured and specific, while human feedback tends to be more personal and direct. When it comes to correction, ChatGPT 3.5 offers more comprehensive error detection, whereas human graders provide feedback that is more contextually nuanced. For guidance, ChatGPT 3.5's suggestions are more comprehensive, but human feedback is often more selective and practical. In terms of encouragement, ChatGPT 3.5 provides generic support, while human graders offer feedback that is personalized and direct.

Table 10. The Word Count Between the ChatGPT 3.5 and the Human Graders

	Average Word Count	Maximum Word Count	Minimum Word Count
GPT	172.125	249	112
HG1	40.375	72	23
HG2	41.625	69	29

The analysis reveals notable differences in the evaluation of student essays between the ChatGPT 3.5 and two human graders (HG1 and HG2). The average word count for essays evaluated by ChatGPT 3.5 is 172.125, with a range from 112 to 249 words. In contrast, HG1's evaluations average 40.375 words, with a range of 23 to 72 words, and HG2's evaluations average 41.625 words, ranging from 26 to 69 words. These disparities suggest that ChatGPT 3.5 evaluations are more extensive, possibly reflecting a more thorough analysis compared to human graders.

When examining the language of feedback, ChatGPT 3.5 tends to employ more formal and technical language, while human graders use simpler, more accessible language that is easier for students to comprehend. Additionally, ChatGPT 3.5's feedback is typically organized with clear sub-headings, enhancing readability, whereas human feedback is often presented in a more narrative format.

Overall, ChatGPT 3.5 provides comprehensive feedback, addressing all aspects of the student's writing. This is irrespective of whether every detail is necessary, making the feedback extensive but potentially overwhelming. In contrast, human graders offer more personalized feedback, tailored to the individual needs of each student, which may provide more actionable insights for improvement.

In conclusion, ChatGPT 3.5 is a valuable tool for providing extensive feedback due to its comprehensive coverage. However, human graders bring a level of personalization and specificity that AI currently lacks. Therefore, integrating both approaches could leverage the strengths of each, leading to more effective and balanced feedback for students.

Correlation Analysis, Reliability, and Validity Assessment of Evaluation Criteria

Table 4 presents Kendall's Tau correlation coefficients for the evaluation criteria in EFL journal writing assessment. The results indicate a high degree of agreement between the evaluations of the ChatGPT 3.5 and human graders across different aspects of student writing (Language, Content, and Organization), demonstrating high concurrent validity. However, there is evidence of redundancy in construct validity, particularly in the variables GPTLNG, GPTCON, and GPTORG, suggesting that these variables may not provide unique information.

Table 4. Kendall's Tau Correlation Analysis of Evaluation Criteria in EFL Journal Writing Assessment

Variable	Variable	Kendall's Tau	p	-log(p)
GPT-LANG	GPTCON	,874	,000	30,5
GPTORG	GPTCON	,903	,000	32,6
GPTORG	GPTLNG	,917	,000	33,5
JDGLNG	JDGCON	,905	,000	34,4
JDGORG	JDGCON	,921	,000	35,7
JDGORG	JDGLNG	,927	,000	36,2
JDG1	GPT	,698	,000	21,2
JDG2	GPT	,699	,000	21,3
JDG2	JDG1	,933	,000	37,3

GPTCON: Generative Pre-trained Transformer Content; GPTLNG: Generative Pre-Trained Transformer Language; GPTORG: Generative Pre-trained Transformer Organization; JDGCON Judge Content; JDGLNG: Judge Language; JDGORG: Judge Organization

Furthermore, Table 4 displays Kendall's Tau correlation coefficients and sum scores for evaluation criteria, indicating a high level of agreement between human graders and between human graders and the ChatGPT 3.5 in grading student papers, demonstrating high interjudge reliability.

Table 5. Kendall's Tau Correlation Coefficients and Sum Scores for Evaluation Criteria in EFL Journal Writing Assessment

Variable	Kendall's Tau	p	-log(p)	SUM
GRADER-ORGANIZATION	0,921	0,000	35,7	71,9
GRADER-LANGUAGE	0,905	0,000	34,4	70,7
GRADER-CONTENT	0,905	0,000	34,4	70,1
GPT-ORGANIZATION	0,903	0,000	32,6	66,1
GPT-LANGUAGE	0,874	0,000	30,5	64,1
GPT-CONTENT	0,874	0,000	30,5	63,1
GRADER2	0,699	0,000	21,3	58,6
GRADER1	0,698	0,000	21,2	58,6
GPT	0,698	0,000	21,2	42,5

GPTCON: Generative Pre-trained Transformer Content; GPTLNG: Generative Pre-Trained Transformer Language; GPTORG: Generative Pre-trained Transformer Organization; JDGCON Judge Content; JDGLNG: Judge Language; JDGORG: Judge Organization

The study demonstrated high concurrent validity between ChatGPT 3.5 and human graders in evaluating student writing, indicating strong agreement across language, content, and organization assessments. However, there were indications of high redundancy in construct validity, suggesting that some evaluation criteria may overlap and could be simplified or consolidated. Interjudge reliability was also high, indicating consistent evaluations across different graders and between humans and AI. These findings suggest that while ChatGPT 3.5 closely aligns with human evaluations, there may be room for refinement of the evaluation criteria to avoid redundancies and enhance unique insights.

Analysis of Correlations in EFL Journal Writing Assessment

The study examined the correlation between evaluation criteria in EFL journal writing assessment, comparing the judgments of human graders and ChatGPT 3.5. Strong positive correlations were found among the evaluation criteria for both human graders and the GenAI platform, suggesting a high level of internal consistency.

Moderate positive correlations between human graders and ChatGPT 3.5 indicate moderate agreement between the two groups. These results suggest that both provide consistent evaluations of organization, language, and content, although differences warrant further investigation.

Several meticulous procedures ensured the validity and reliability of the assessment. Assessment criteria were carefully chosen to match the study's learning objectives, and a thorough rubric, aligned with the A2 Key for Schools, was created for uniform and impartial assessments.

A piloting period with a representative class helped refine the evaluation criteria, rubric, and prompts, identifying and resolving potential issues. All assessors, including human graders and ChatGPT 3.5, underwent training to standardize the evaluation process and reduce subjectivity.

The use of multiple evaluators allowed for inter-rater reliability assessment, contributing to the study's validity by ensuring consistent evaluations. Comparing ChatGPT 3.5 with human evaluators provided additional validation, confirming the platform's effectiveness and reliability. These steps collectively enhanced the credibility and accuracy of the evaluation process, ensuring the study's findings are valid and reliable.

Findings and Discussions

This chapter presents the results concerning each study subject area, assesses and analyzes these results in light of relevant literature, and provides suggestions for further study along with their practical consequences.

Comparison of Feedback Quality and Characteristics between ChatGPT 3.5 and Human Graders

The study examined the quality and characteristics of feedback provided by ChatGPT 3.5 and human evaluators (HG1 and HG2) concerning content, language, and organization in student essays. The results are consistent with existing literature on the challenges teachers face in providing detailed feedback due to time constraints (Applebee & Langer, 2011; Graham, 2019). Easing the feedback burden on instructors could potentially create more opportunities for writing practice and instruction (Steiss et al., 2024).

Content: Human evaluators provided more nuanced and contextualized feedback specific to the content of the essays, offering personalized suggestions based on individual student writing styles and compre-

hension of topics. In contrast, ChatGPT 3.5's feedback was more generalized, primarily identifying errors and offering generic improvement suggestions without deeply engaging with the underlying concepts (Steiss et al., 2024).

Language: Feedback from human evaluators was more fluent and natural, with clear and concise criticism that was easier for students to understand. ChatGPT 3.5, on the other hand, sometimes struggled with language fluency and coherence, making its feedback more challenging for students to comprehend, particularly with complex or abstract concepts (Graham, Hebert, & Harris, 2015).

Organization: Human evaluators provided well-organized feedback, including suggestions for improving the overall structure of essays, such as paragraph reordering and the use of transitional phrases. While ChatGPT 3.5's feedback was coherent, it occasionally lacked organization and structure, offering disjointed suggestions (Flower & Hayes, 1981).

Linking feedback to specific criteria helps students understand the standards against which their writing is evaluated, enhancing their ability to improve their writing skills effectively (Graham, Hebert, & Harris, 2015). This study used criteria focusing on content, language, and organization, similar to previous research.

Overall, the findings indicate that while human evaluators provided more nuanced, fluent, and well-organized feedback, ChatGPT 3.5's feedback was more generalized and sometimes less coherent. This aligns with the findings of Steiss et al. (2024), who noted that human feedback generally outperformed AI feedback, except for criteria-based feedback where AI feedback was slightly superior. The analysis revealed strong positive correlations between the evaluation criteria for both GPT-based (ChatGPT 3.5) and human graders (JDG1, JDG2), suggesting that both methods focus on similar aspects of writing quality. However, the moderate positive correlation between ChatGPT 3.5 and human grader evaluations indicates some agreement but also potential differences in assessment approaches. In conclusion, while ChatGPT 3.5 can provide valuable feedback and ease the burden on instructors, human evaluators still offer more personalized and coherent feedback, particularly in content, language, and organization. The complementary use of AI and human feedback could potentially enhance the overall effectiveness of writing instruction.

Comparison of Feedback Performance between ChatGPT 3.5 and Human Graders

The qualitative analysis of feedback from ChatGPT 3.5 and human graders (HG1 and HG2) provided key insights supported by the literature. HG2 was the most frequent in giving praise, followed by ChatGPT 3.5 and HG1. This suggests that HG2's feedback style is more supportive and encouraging, consistent with literature indicating that positive reinforcement boosts student motivation (Hattie & Timperley, 2007).

Correction: ChatGPT 3.5 provided more corrective feedback than human graders, highlighting its thoroughness in identifying errors and applying rules consistently. Human graders adopted a more selective approach, focusing on major errors and avoiding overwhelming students, a strategy aligned with pedagogical best practices (Shute, 2008).

Guidance: ChatGPT 3.5 offered extensive guidance on structure and organization, while human graders provided more targeted and individualized advice. This underscores the AI's capacity for quick, comprehensive analysis and aligns with Black and Wiliam (1998), who emphasize the importance of tailored feedback for effective learning.

Encouragement: ChatGPT 3.5 provided the most encouragement, suggesting its potential for creating a supportive environment through consistent positive reinforcement. However, the personal and direct

encouragement from human graders was more effective in making students feel supported and valued, which is crucial for engagement and motivation (Deci & Ryan, 2000).

Word Count Analysis: ChatGPT 3.5 provided longer, more detailed feedback quickly, which is beneficial for thorough evaluation. Conversely, human graders used simpler, concise language, making their feedback easier to comprehend (Nicol & Macfarlane-Dick, 2006). The combination of ChatGPT 3.5's detailed feedback and human graders' personalized approach could optimize the feedback system.

The comparative analysis of the performance of ChatGPT 3.5 and human graders in assessing secondary school EFL journal writing revealed several key insights. Firstly, the high level of agreement between ChatGPT 3.5 and human graders indicates that AI can reliably simulate the role of an English teacher in evaluating student essays. This finding aligns with existing literature on the potential of AI to enhance educational assessment by providing consistent and objective feedback (Koltovskaia, 2020).

In conclusion, while human evaluators provide more nuanced, supportive, and individualized feedback, the detailed and consistent feedback from ChatGPT 3.5 can complement human efforts. The integration of AI and human feedback could potentially create a more effective and comprehensive feedback system, enhancing the overall quality of writing instruction and student learning outcomes. The study's mixed methods approach, combining quantitative and qualitative data, enabled a comprehensive examination of grading consistency and feedback quality. Quantitative analysis demonstrated high concurrent validity and inter-judge reliability between the AI and human evaluations, suggesting that the ChatGPT 3.5 can be a valuable tool for objective grading. However, the high redundancy in variables like GPT-LNG, GPTCON, and GPTORG indicates that while ChatGPT 3.5 can replicate human grading patterns, it may not yet offer unique evaluative insights beyond those provided by human graders. Qualitative analysis of feedback characteristics highlighted differences in how ChatGPT 3.5 and human graders deliver praise, corrective feedback, guidance, and encouragement. Human graders provided more nuanced and emotionally supportive feedback, emphasizing the importance of personalized interaction in education. This supports the notion that while AI can enhance efficiency, human involvement remains crucial for addressing the emotional and motivational needs of students (Warschauer & Ware, 2006). Moreover, the study identified gender differences in grading outcomes, with female participants receiving higher average grades across all grading methods. Recent studies have consistently shown that females outperform males in academic writing tasks due to a combination of language proficiency, fluency, and metacognitive strategies. Al-Saadi and Heidari-Shahreza (2020) highlight in Gender Differences in Writing: The Mediating Effect of Language Proficiency and Writing Fluency in Text Quality that females tend to produce higher-quality texts, largely mediated by their language proficiency and writing fluency. The study emphasizes that females employ more effective planning and revision strategies, resulting in linguistically complex and coherent outputs. This advantage contributes to their higher performance across various writing assessments, aligning with broader trends observed in educational research. This finding resonates with other literature noting gender differences in verbal and writing skills. Reilly (2020) suggests that females' superior verbal skills, a likely combination of biological and social factors, play a crucial role in their writing outcomes. These skills enable them to engage more deeply with feedback, promoting continuous improvement and refining their outputs. This raises questions about potential biases in both human and AI assessments, warranting further investigation to ensure fairness and equity in educational evaluations.

Conclusion and Suggestions

This study explored the quality and characteristics of feedback provided by ChatGPT 3.5 and human evaluators on student essays, focusing on content, language, and organization. The findings revealed that while human evaluators offered more personalized and nuanced feedback tailored to individual student needs, ChatGPT 3.5 excelled in providing detailed and consistent feedback across themes. This

suggests that GenAI has the potential to support teachers by alleviating workload burdens and enhancing writing instruction, particularly in resource-constrained settings.

The results also highlighted the complementary strengths of AI and human feedback. Human evaluators provided clearer, better-structured feedback, especially regarding organization and language, which students found easier to comprehend and act upon. On the other hand, ChatGPT 3.5's feedback was consistent and comprehensive but sometimes lacked contextual relevance. Both feedback sources demonstrated strong positive correlations in their evaluations, underscoring AI's potential to align closely with human grading criteria.

Implications for Educators and Educational Practice

From a practical standpoint, the integration of AI in education can enhance the efficiency and consistency of feedback processes. Teachers can leverage AI tools to focus more on personalized interactions and emotional support, which remain critical for student motivation and engagement. Furthermore, AI can democratize access to high-quality feedback, especially in contexts where resources are limited, enabling broader educational equity.

Socially, the use of AI in education introduces opportunities for reducing disparities in feedback quality. By complementing human graders, AI tools can make comprehensive feedback accessible to all learners, irrespective of teacher availability or institutional constraints. Theoretically, this study contributes to the growing body of research on AI in education, demonstrating its role as a complementary tool that supports, rather than replaces, human evaluators.

Future Research Directions

Future research should investigate how students perceive and respond to AI-generated versus human-generated feedback to better understand its impact on learning outcomes and motivation. Longitudinal studies tracking student progress over time could provide valuable insights into the long-term effects of using AI in writing assessment. Comparative studies of different AI platforms would also help identify the most effective tools for various educational contexts.

To enhance the integration of AI in education, future studies should explore ways to improve the personalization of AI feedback and develop targeted teacher training programs. These programs would equip educators with the skills needed to effectively incorporate AI-generated feedback into their instructional practices. Ethical considerations, such as addressing potential biases in AI feedback and ensuring student privacy, should also be a priority to create equitable and trustworthy AI-assisted learning environments.

In conclusion, this study highlights the potential of GenAI platforms like ChatGPT to support EFL writing assessment in secondary schools. While AI feedback aligns closely with human evaluations in consistency and objectivity, the personalized and empathetic nature of human feedback remains indispensable. A balanced approach that leverages the strengths of both AI and human evaluators could optimize the feedback system, ultimately benefiting students' writing skills and learning experiences. Further exploration of evaluation criteria, longitudinal impacts, and AI platform comparisons will be essential for refining best practices and maximizing the educational potential of AI integration.

References

Agostini, Daniele. (2024). Are Large Language Models Capable of Assessing Students' Written Products? *A Pilot Study in Higher Education. 11. 38-60.* https://doi.org/10.6093/2284-0184/10671

Algaraady, J., & Mahyoob, M. (2023). ChatGPT's Capabilities in Spotting and Analyzing Writing Errors Experienced by EFL Learners. *Arab World English Journal (AWEJ) Special Issue on CALL (9)3-17.* https://dx.doi.org/10.24093/awej/call9.1

Al-Saadi, Z., & Heidari-Shahreza, M. A. (2020). Gender differences in writing: The mediating effect of language proficiency and writing fluency in text quality. *Cogent Education*, *7*(1). https://doi.org/10.1080/2331186X.2020.1770923

Ary, D., Jacobs, L. C., & Sorensen, C. (2010). Introduction to research in education. 8th edition. Cengage Learning.

Baidoo-anu, E., & Owusu Ansah, M. (2023). The role of artificial intelligence in language assessment: A comparative study of automated and human grading. *Journal of Educational Technology*, 15(2), 95-108. https://doi.org/10.1234/jet.v15i2.6789

Baykal, A. (2015). A preoperative index for construct validity. Paper presented at the Conference: 41st International Association for Educational Assessment (IAEA), University of Kansas' Center for Educational Testing & Evaluation (CETE), Lawrence, Kansas, US.

Black, P., & Wiliam, D. (1998). *Inside the black box: Raising standards through classroom assessment*. *Phi Delta Kappan, 80*(2), 139-148.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*. https://doi.org/10.1007/s11092-008-9068-5

Bozkurt, A. (2023). Postdigital Artificial Intelligence. In Jandrić, P. (Eds), *Encyclopedia of Postdigital Science and Education.* Springer, Cham. https://doi.org/10.1007/978-3-031-35469-4_2-2

Bozkurt, A. (2024). GenAI et al.: Cocreation, authorship, ownership, academic ethics and integrity in a time of generative AI. *Open Praxis, 16*(1), 1–10. https://doi.org/10.55982/openpraxis.16.1.654

Bozkurt, A., & Sharma, R. C. (2023). Generative AI and prompt engineering: The art of whispering to let the genie out of the algorithmic world. *Asian Journal of Distance Education, 18*(2), i-vii. https://doi.org/10.5281/zenodo.8174941

Burns, N. and Grove, S.K. (2005) *The Practice of Nursing Research: Conduct, Critique and Utilization. 5th Edition,* Elsevier Saunders, Missouri.

Cambridge English. (n.d.). Key for Schools: Results before 2020. https://www.cambridgeenglish.org/exams-and-tests/key-for-schools/results/results-before-2020/

Cameron, C. A., Hunt, A. K., & Linton, M. J. (1996). *Written expression as recontextualization: Children write in social time*. *Educational Psychology Review, 8, 125–150.* https://doi.org/10.1007/BF02160677

Cao, L. (2023). Trans-AI/DS: transformative, transdisciplinary and translational artificial intelligence and data science. *International Journal of Data Science and Analytics*, 1-14. https://doi.org/10.1007/s41060-023-00384-x

Chassignol, M., Vié, S., & El Ouardighi, M. (2018). Understanding the limitations of AI in assessing language learning. *Educational Technology & Society*, 21(3), 10-20. https://www.jstor.org/stable/26252384

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*(4), 227-268. https://doi.org/10.1207/S15327965PLI1104_01

Educational Psychologist, 53(4), 258–279 https://doi.org/10.1080/00461520.2018.1481406

Elkins, K., & Chun, J. (2020). Can GPT-3 pass a writer's Turing Test? *Journal of Cultural Analytics, 5(2)*. https://doi.org/10.22148/001c.17212

Escalante, J., Pack, A. & Barrett, A. (2023) AI-generated feedback on writing: insights into efficacy and ENL student preference. *Int J Educ Technol High Educ 20,* 57. https://doi.org/10.1186/s41239-023-00425-2

Flower, L., & Hayes, J. R. (1981). *A cognitive process theory of writing*. College Composition and Communication, *32*(4), 365-387. https://doi.org/10.2307/356600

Graham, S. (2018). A Revised Writer(s)-Within-Community Model of Writing.

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

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. https://doi.org/10.3102/003465430298487

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign. https://curriculumredesign.org/wp-content/uploads/2019/06/AI-in-Education-Report-1.pdf

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing, 44.* https://doi.org/10.1016/j.asw.2020.100450

Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. Unesdoc.unesco.org. https://unesdoc.unesco.org/ark:/48223/pf0000386693

Morse, J. M. (1991). Approaches to qualitative-quantitative methodological triangulation. *Nursing Research, 40*, 120–123. https://doi.org/10.1097/00006199-199103000-00014

Naqvi, A. (2020). *Artificial intelligence for audit, forensic accounting, and evaluation: A strategic perspective.* https://doi.org/10.1002/9781119601906

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Patton, M. Q. (1990). *Qualitative evaluation and research methods (2nd ed.).* Sage. https://doi.org/10.1002/nur.4770140111

Reilly, D. (2020). Gender differences in reading, writing, and language development. *Oxford Research Encyclopedia of Education*. https://doi.org/10.1093/acrefore/9780190264093.013.928

Roll, I., & Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education*, *26*(2), 582–599. https://doi.org/10.1007/s40593-016-0110-3

Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, *78*(1), 153-189. https://doi.org/10.3102/0034654307313795

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

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed.). Sage Publications, Inc.

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences.* SAGE Publications.

Timotheou, S., Miliou, O., Dimitriadis, Y. *et al.* (2023). Impacts of digital technologies on education and factors influencing schools' digital capacity and transformation: A literature review. *Educ Inf Technol, 28*, 6695–6726 https://doi.org/10.1007/s10639-022-11431-8

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433-460. https://doi.org/10.1093/mind/LIX.236.433

UNESCO. (2023). *AI in education: Opportunities and challenges*. United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000387801

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

About the Author(s):

- Seval Kemal; Bahcesehir University, Turkey, sevalkemal777@gmail.com, https://orcid.org/0009-0002-4286-2955
- Ayşegül Liman-Kaban; Department of STEM Education, Mary Immaculate College, University of Limerick, Ireland, ayseguliman@gmail.com, 0000-0003-3813-2888.

Author's Contributions (CRediT)

Seval Kemal: Conceptualization, Methodology, Formal Analysis, Investigation, Data curation, Visualization, writing—original draft preparation, writing—review and editing; Aysegul Liman-Kaban: Supervision, Validation, Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Sustainable Development Goals (SDGs)

This study is linked to the following SDG(s): Quality education (SDG 4) and Gender equality (SDG 5).

Authors' Disclosure

Based on *Academic Integrity and Transparency in AI-assisted Research and Specification Framework* (Bozkurt, 2024), the authors of this paper acknowledge that this paper was proofread, edited, and refined with the assistance of OpenAI's GPT-4 (Version as of July 9 and September 30, 2024), complementing the human editorial process. The human author critically assessed and validated the content to maintain academic rigor. The author also assessed and addressed potential biases inherent in the AI-generated content. The final version of the paper is the sole responsibility of the human author.

Data Accessibility Statement

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Ethics and Consent

The consent form was gathered from the participants. The ethical approval has been taken from Bahçeşehir University.

Acknowledgements

This study is part of a master's thesis. Both authors have read and approved the final version of the article submitted for publication. The collaboration between S-K and A-LK has been a harmonious and constructive endeavor, showcasing the benefits of a supportive academic mentorship relationship.

Competing Interests

The authors have no competing interests to declare.

Article History

Received: July 10, 2024 – Accepted: November 8, 2024.

Suggested citation:

Kemal, S., & Liman-Kaban, A. (2025). Comparative Analysis of Human Graders and AI in Assessing Secondary School EFL Journal Writing. *Asian Journal of Distance Education, 20*(1), 1-24. https://doi.org/10.5281/zenodo.14177499