Studies in English Language Teaching
ISSN 2372-9740 (Print) ISSN 2329-311X (Online)
Vol. 14, No. 3, 2026
www.scholink.org/ojs/index.php/selt

Original Paper

A Comparative Analysis of AI-Generated Feedback Tools on EFL Students' Writing in Higher Education

Mona Abdullah Alzahrani[1*]

[1] Taif University, Saudi Arabia

[*] Corresponding Author, E-mail: m.alzahrane@tu.edu.sa

Received: July 23, 2026 Accepted: August 24, 2026 Online Published: August 28, 2026

doi:10.22158/selt.v14n3p146 URL: http://dx.doi.org/10.22158/selt.v14n3p146

Abstract

This research compared the feedback of three generative artificial intelligence (AI) tools, ChatGPT, Google Gemini, and Microsoft Copilot, on EFL students' writing. It adopted a mixed-methods design to evaluate the writing of 35 medical students who enrolled in an Intensive English for Academic Purposes (IEAP) course at Taif University during 2025-2026. The AI-generated feedback was coded into five categories: grammar, vocabulary, cohesion, sentence restructuring, and mechanics. Quantitative data were analyzed through descriptive statistics and one-way repeated measures ANOVA, while qualitative data were examined through thematic analysis of feedback style and depth. The findings revealed that grammar was observed to be the most common feedback type in all three tools. Through a one-way repeated measures ANOVA test, there were no significant differences in the feedback on grammar and cohesion in all three tools. There were significant differences in the feedback on vocabulary, sentence restructuring, and mechanics. MS Copilot had a significantly higher number of vocabulary and sentence restructuring feedback than ChatGPT and Gemini, while ChatGPT had a significantly higher number of mechanics feedback. The qualitative analysis showed different types of feedback where ChatGPT worked as a corrective feedback tool, Gemini worked on instructional methods combining corrections with explanations and suggestions, and MS Copilot used text improvements through sentence reformulation. The research will contribute to the existing literature regarding AI-assisted writing by demonstrating that AI tools differ in the feedback they provide and in their pedagogical focus, style, and depth.

Keywords

Artificial intelligence, AI-generated tools, automated feedback, EFL writing, higher education

Introduction

In recent times, learning, teaching, and assessment activities have been shaped by the incorporation of

artificial intelligence (AI) technologies in educational institutions (Kurtz et al., 2024). The impact of AI technology is particularly pronounced in EFL education (Farazouli et al., 2024; Guo & Wang, 2023; Link et al., 2022). One of the areas where AI technologies are extensively applied is the assessment of writing skills. Indeed, writing is one of the fundamental skills that is necessary for academic and professional success (Labadze et al., 2023; Nguyen et al., 2024). According to Zhao et al. (2024), the digitalization of writing is a significant shift in the writing process that has a major impact on the process of writing. Hence, today's learners frequently use writing assistants that are based on AI technologies to develop, edit, and improve their written texts. Such AI-based technologies provide new possibilities of enhancing writing through automation of feedback on several aspects of writing, such as grammar, vocabulary, and organization (Peláez-Sánchez et al., 2024).

It is widely accepted that feedback is an important element of second language writing instruction. By helping students recognize their language-use errors, develop their ideas, and master high-level writing skills such as coherence and organization, it plays an important role in improving their writing (Shi, 2021; Wang, 2021). Good feedback motivates students to involve themselves in the writing process and rewrite, not only to fix their mistakes. In previous research, it was found that the focus of different types of feedback may vary; while some of them may emphasize the global features of writing, like content and structure, others may emphasize the local aspects of writing, like grammar and mechanics (Cheng & Zhang, 2021). With the emergence of AI-generated feedback, the issue becomes even more complicated and raises questions about whether computer-based feedback is comparable to human feedback. With the help of fast, reliable, and efficient feedback from automated writing evaluation (AWE) tools, students can make better edits to their writing. It is also known that the AWE tools can improve writing results, especially in combination with opportunities for rewriting (Wei et al., 2023).

The utility of AI tools for generating feedback regarding students' writing assignments has been investigated recently. It was found that feedback generated by artificial intelligence could enhance the involvement of learners and their independence, as well as improve grammar, clarity, and quality of writing (Polakova & Ivenz, 2024). Besides, it appears that recent research shows certain differences in the functionality of various AI technologies. Although generative AI tools can generally be efficient in organizing text and improving their structure, some particular AI programs have been found to be useful in identifying writing errors and helping writers revise their texts (Al Sofi, 2026; Wei et al., 2023). However, the comparison of AI-generated feedback and feedback provided by professionals showed that there were certain aspects in which AI technology was not as efficient as professional teachers (Steiss et al., 2024).

Although great progress has been made in developing AI tools, there still lacks comprehensive research comparing several different AI tools and their feedback. Most research is devoted either to certain tools or aspects of their functioning. Little emphasis has been placed on how feedback is presented (e.g., direct correction, suggestions, or rewriting) and how deeply it engages with the writing process. Addressing this gap is critical for assessing the pedagogical benefit of AI tools and their role in improving both

surface-level correctness and higher-order writing skills. Therefore, the present study aims to address this gap by conducting a systematic comparison of three widely used AI writing tools—ChatGPT, Google Gemini, and Microsoft Copilot. To reach this goal, the following three research questions were proposed to guide this study:

1- What types of feedback do ChatGPT, Microsoft Copilot, and Gemini provide on students' writing?

2- Are there statistically significant differences among the three AI tools—ChatGPT, Gemini, and Microsoft Copilot—in the types of feedback provided?

3- How do AI tools differ in terms of feedback style and depth?

Literature Review

Feedback on EFL writing

It has been seen that feedback is considered a basic part of EFL writing learning as it helps students to find mistakes, revise their ideas, and develop linguistically accurate and better writing (Olsen & Hunnes, 2023). Feedback assists in developing different aspects of writing, such as grammatical aspects, lexical aspects, cohesion, organization, and other aspects of writing, while encouraging revision as well as reflective learning (Shi, 2021; Wang, 2021). Effective feedback not only involves error correction but also helps the learner to learn the reason behind making certain changes.

Nevertheless, the effectiveness of feedback can be influenced not only by the substance of the feedback itself but also by how learners respond to it. Engagement of learners with feedback can be viewed as a multidimensional process encompassing affective, behavioral, and cognitive aspects of learners' responses. The affective component of learners' engagement is defined as their emotional reaction towards feedback. The behavioral component can be considered as learners' behavior in revising their texts, while the cognitive one represents deeper processing of the feedback by learners (Teng & Huang, 2025). Research indicates that clear and actionable feedback enhances learners' engagement and, consequently, their writing performance (Shi, 2021; Teng & Huang, 2025).

Although teacher feedback is beneficial, it could be expensive and inconsistent in many ways, particularly in the case of big groups, while peer feedback may not help in correcting the structural errors, that is, the organization and content errors. This kind of problem led to the idea that there were other methods of making improvements in the process, and one of them was technology-based feedback (Guo & Wang, 2024; Teng & Huang, 2025).

Automated writing evaluation and generative feedback

The use of AI-powered educational tools may offer customized teaching opportunities, meet the needs of each learner, and even evaluate the task (Zhai, 2022). Therefore, the rapid growth of artificial intelligence (AI) and Automated Writing Evaluation (AWE) has drastically changed the feedback methods in EFL writing. They have made it possible to deliver feedback immediately, at large volumes, and tailored to learners' unique needs, solving most of the problems associated with conventional feedback approaches (Bond et al., 2024). Such tools could assess learners' writing using multiple criteria, including content,

structure, grammar, and coherence (Gill et al., 2024; Wei et al., 2023).

The existing literature on AWE tools like Write & Improve shows that AI feedback is equally effective to the teacher's feedback in helping students improve their performance in writing in terms of its coherence, lexical resource, and overall quality (Bodaubekov et al., 2025). On the other hand, many researchers have found that the feedback provided by automatic systems helps to increase the accuracy of writing, learners' autonomy, and interaction in L2 settings, especially when it is incorporated into classroom instruction. However, there is evidence that AI feedback tools are especially useful in terms of addressing surface-level features of writing, such as grammar and spelling (Guo & Wang, 2024; McCarthy et al., 2022).

However, with the advent of generative AI tools like ChatGPT, the capabilities of conventional AWE applications have been further expanded owing to the fact that the feedback is now more flexible, adaptive, and interactive. In contrast to conventional feedback provided using pre-made templates, generative AI provides adaptive and personalized feedback. Therefore, it has immense potential in supporting different phases of writing from generation to revision and editing (Peláez-Sánchez et al., 2024; Teng & Huang, 2025). Literature suggests that ChatGPT improves learners' motivation, self-efficacy, and engagement, besides helping with their writing skills like organization and vocabulary use (Li et al., 2024; Teng & Huang, 2025).

In empirical research, it was found that generative AI tools deliver more comprehensive and well-balanced feedback as opposed to the feedback from teachers. ChatGPT, for instance, was found to deliver considerably more feedback than human teachers do and allocate attention evenly between content, organization, and language issues (Guo & Wang, 2024). At the same time, feedback generated by an AI tool is characterized not only by quantity but also by quality, since AI-generated feedback usually includes different types of feedback, such as direct correction, suggestion, explanation, or summary. Nevertheless, certain differences in feedback style also occur because the feedback delivered by teachers is more informative and based on questions (Guo & Wang, 2024).

At present, there are several gaps in the literature about AI feedback in L2 writing that need further investigation. Firstly, most existing studies concentrate on individual AI programs such as ChatGPT or AWE systems but do not provide comparative analyses of various AI technologies, even though there is evidence of differences in feedback provided by different systems. Secondly, while recent studies have concentrated on learners' interactions with AI feedback from the perspective of affective, behavioral, and cognitive aspects, the role of certain characteristics of AI feedback in relation to writing development and learning outcomes is still unknown. Finally, given that the findings about the effectiveness of human and artificial intelligence feedback are controversial, a more comprehensive approach to studying the problem should be elaborated. The listed problems should be addressed to understand the use of AI in second language writing.

Methodology

Research Design

In this study, a mixed-methods comparative research design has been used to analyze the feedback generated by the three different generative AI tools – ChatGPT, Google Gemini, and Microsoft Copilot. A mixed-methods research design is adopted in the current study as it allows for an analysis of the feedback generated by AI tools in terms of quantitative and qualitative approaches. In the quantitative approach, the frequency, percentage, and mean scores of the different types of feedback generated by the different tools have been analyzed, while in the qualitative approach, differences between the focus and depth of the feedback have been explored. Such a combination of methods allowed us to receive a more comprehensive idea about the educational value of AI-generated feedback and its patterns, both quantitative and qualitative.

Participants

This study consisted of 35 students registered in the Faculty of Medicine at Taif University, Saudi Arabia, during the fall semester of the academic year 2025–2026. These were first-year students of medicine, having an English language proficiency level varying between lower-intermediate and intermediate. All students were registered for a mandatory Intensive English for Academic Purposes (IEAP) course requiring them to write regularly on eight different topics. A total of 35 writing samples were collected and used as the dataset for this study. To ensure consistency, the same writing samples were submitted to three AI tools: ChatGPT, Google Gemini, and Microsoft Copilot, with identical prompts, enabling a direct comparison of the feedback generated by each system.

Instruments

The primary research instruments used in this study were a corpus of student writing samples and a feedback analysis framework (see Appendix 1). The corpus consisted of 35 writing samples. These samples were submitted to three generative AI tools—ChatGPT, Google Gemini, and Microsoft Copilot—using the same prompt "Review my one-sided opinion paragraph and give feedback on clarity, structure, grammar, spelling, punctuation, and vocabulary." to ensure consistency in feedback generation. The feedback generated by each tool served as the primary data source for analysis.

To analyze the feedback systematically, a coding scheme was developed based on categories commonly used in studies of written corrective feedback, automated writing evaluation, and AI-assisted writing feedback (Cheng & Zhang, 2021; Guo & Wang, 2024; McCarthy et al., 2022; Wei et al., 2023). It categorized feedback into five categories: grammar, vocabulary, cohesion, sentence restructuring, and mechanics. These categories were selected because they represent the major linguistic and discourse-level dimensions commonly addressed in writing assessment and feedback research.

Moreover, feedback style was analyzed using the coding scheme, whereby feedback can be given in the form of direct correction, suggestion, and rewriting, while feedback depth was analyzed in terms of being low, medium, or high depending on the explanation that is given (see Appendix 1).

Data Collection

Before collecting the data, students signed the consent form that explained their rights, which involved voluntary participation, withdrawing anytime they wanted, and ensuring anonymity and confidentiality. Each student worked individually on the writing task for approximately 45 minutes. Firstly, students were required to write a one-sided opinion paragraph with an introduction with a clear thesis statement, body sentences with supporting details, and a concluding sentence. Subsequently, the writing samples were inserted into ChatGPT, Google Gemini, and Microsoft Copilot by every student under the same prompt, asking for feedback on clarity, organization, grammar, spelling, punctuation, and vocabulary. Finally, all the feedback comments provided by the three AI programs were recorded and used to form the dataset. Every feedback comment was coded using a predetermined framework derived from the literature on writing feedback and automated writing evaluation.

Data Analysis

Data analysis consisted of both quantitative and qualitative procedures. Quantitative analysis included the distribution of feedback types that are given by ChatGPT, Gemini, and MS Copilot. For each type of feedback, descriptive statistics such as frequencies, percentages, means, and standard deviation were obtained. To find out whether there are any differences between the AI tools in their ability to give feedback on the issues of grammar, vocabulary, cohesion, sentence restructuring, and mechanics, inferential statistical analyses were performed using one-way repeated measures ANOVA. One-way repeated measures ANOVA was chosen since each of the essays was analyzed by all three AI tools. The Greenhouse-Geisser correction was employed when there is a violation of the assumption of sphericity

Qualitative analysis, on the other hand, included the study of the nature of the feedback provided by each AI tool. Thematic analysis was utilized in this research to examine feedback style and depth. Feedback comments were coded according to the predefined styles of direct correction, suggestion, rewriting, list format, and paragraph feedback. In addition, feedback was analyzed for depth and classified as low, medium, or high. Low means corrections only. Medium means corrections accompanied by limited suggestions for improvement. High means corrections accompanied by explanations and rewriting. Furthermore, AI tools were categorized based on their main function. Through the integration of quantitative and qualitative findings, cross-validation was enabled, and the study provided a comprehensive comparison of ChatGPT, Google Gemini, and Microsoft Copilot as AI-powered feedback providers in the context of EFL writing.

Findings

RQ1. What types of feedback do ChatGPT, Microsoft Copilot, and Gemini provide on students' writing?

This study analyzed feedback from three AI writing tools—ChatGPT, Gemini, and Microsoft Copilot—across 35 student writing samples. To answer this question, feedback was coded into five categories: grammar, vocabulary, cohesion, sentence restructuring, and mechanics. The total frequencies for each category are presented in Table 1.

Table 1. Frequencies and Percentages of Feedback Types by AI Tool

Feedback Type	ChatGPT		Gemini		MS Copilot	
	F	%	F	%	F	%
Grammar	166	48.7	147	48.4	144	34.7
Vocabulary	57	16.7	58	19.1	70	16.9
Cohesion	41	12.0	36	11.8	42	10.1
Restructuring	22	6.5	23	7.6	62	34.7
Mechanics	55	16.1	38	12.5	45	10.8

Table 1 displays the distribution of feedback categories in percentages provided by ChatGPT, Google Gemini, and Microsoft Copilot regarding the writings of the students. It can be seen that the highest category of feedback that was provided by ChatGPT and Google Gemini was grammar, which accounted for 48.7% and 48.4%, respectively. Such results demonstrate that these two tools pay special attention to grammar and language accuracy. Meanwhile, grammar is one of the top two feedback types offered by MS Copilot (34.7%). However, it is noticeable that the percentage of grammar feedback from MS Copilot is much lower compared to ChatGPT and Google Gemini.

One of the significant distinctions in the analysis of the three tools lies in the area of sentence restructuring. Despite the small amount of restructuring feedback given by ChatGPT (6.5%) and Google Gemini (7.6%), MS Copilot gives a larger amount of restructuring feedback (34.7%). Thus, MS Copilot gives the second highest percentage of the most prominent types of feedback in its results. It means that MS Copilot concentrated more on sentence restructuring and organization.

In terms of vocabulary feedback, there were similarities between the three AI tools. In terms of providing vocabulary feedback, Gemini was the top contributor at 19.1%, while MS Copilot was second at 16.9%, and ChatGPT came third at 16.7%. The findings indicate that vocabulary improvement was not the top priority for all three AI tools.

Cohesion-related feedback was relatively low among the three AI tools and consisted of 12.0% of ChatGPT's feedback, 11.8% of Gemini's feedback, and 10.1% of MS Copilot's feedback. This implies that the three tools were relatively focused on the higher level of writing rather than lower-level features like cohesion.

Like the previous categories of feedback, mechanics feedback (spelling, punctuation, capitalization) accounted for a relatively moderate portion of feedback provided by the AI tools. ChatGPT was the highest in giving mechanics feedback at 16.1%, followed by Gemini at 12.5% and MS Copilot at 10.8%. This shows that ChatGPT allocated relatively great attention to punctuation, spelling, capitalization, and formatting issues.

In conclusion, the analysis of results showed that each of the AI tools has its own feedback pattern. Both ChatGPT and Gemini mostly provided grammar-related feedback, while MS Copilot paid equal attention

to grammar and restructured sentences. While all three tools considered such aspects of writing as vocabulary, cohesion, and mechanics, those were much less frequent than grammar and restructuring. This shows that the three tools are different in the types of feedback they provide.

To better understand the patterns found in the frequency analysis, the descriptive statistics for each category were calculated. Table 2 below presents the means and standard deviations for the number of feedback cases for each category of feedback.

Table 2. Means (M) and Standard Deviations (SD) of Feedback Types by AI Tool

Feedback Type	ChatGPT M (SD)	Gemini M (SD)	MS Copilot M (SD)
Grammar	4.74 (2.582)	4.20 (2.260)	4.11 (1.641)
Vocabulary	1.63 (0.731)	1.66 (0.684)	2.00 (0.542)
Cohesion	1.17 (0.514)	1.03 (0.618)	1.20 (0.473)
Restructuring	0.63 (0.690)	0.66 (0.639)	1.77 (1.911)
Mechanics	1.57 (1.092)	1.09 (0.981)	1.29 (1.017)

The results of the analysis of the frequency of feedback for the five feedback categories delivered by ChatGPT, Google Gemini, and Microsoft Copilot are given in Table 2, where the mean scores and the standard deviations for each type are shown. Based on the results of the table, it is clear that among the three AI programs, the type of feedback provided that was most frequently received was grammar feedback. ChatGPT had the highest mean value of the number of grammar feedbacks (M= 4.74, SD = 2.582). Then there was Gemini (M= 4.20, SD = 2.260) followed by MS Copilot (M = 4.11, SD = 1.641). Therefore, all three tools provided feedback on correcting grammatical errors.

Regarding vocabulary feedback, its frequency did not differ greatly between the three AI tools. MS Copilot provided the largest average frequency (M = 2.00, SD = 0.542), then came Gemini (M = 1.66, SD = 0.684) and ChatGPT (M = 1.63, SD = 0.731). The relatively small standard deviations indicate that vocabulary feedback was provided consistently across the writing samples.

As far as cohesion feedback is concerned, all three models have produced low mean frequency values. MS Copilot has the highest mean value (M = 1.20, SD = 0.473), followed by ChatGPT (M = 1.17, SD = 0.514), and the lowest value is from Gemini (M = 1.03, SD = 0.618). These numbers suggest that the discourse-level features like coherence and transitions have not been paid much attention as compared to grammar and vocabulary.

A significant disparity exists in the number of restructuring comments provided by each model.

Moreover, it can be seen from Table 2 that ChatGPT (M = 0.63, SD = 0.690) and Gemini (M = 0.66, SD = 0.639) have produced very few comments in this regard, while MS Copilot has produced many more (M = 1.77, SD = 1.911). This means that MS Copilot is more concerned with proposing changes to the structural and expression features of the sentences, implying a stronger emphasis on clarity and

readability. The higher value of standard deviation suggests that there is a higher degree of variability in MS Copilot's restructuring suggestions.

The tool that provided the highest mean frequency in the mechanics category was ChatGPT (M = 1.57, SD = 1.092), MS Copilot (M = 1.29, SD = 1.017), and Gemini (M = 1.09, SD = 0.981). This indicates that ChatGPT concentrated on such aspects of writing as punctuation, capitalization, spelling, and formatting.

In general, the results of the descriptive analysis show that the patterns of feedback from the three AI-based tools differ. ChatGPT provided the most frequent feedback in grammar and mechanics categories, Gemini showed somewhat similar patterns with fewer numbers, and MS Copilot focused mostly on sentence restructuring. Thus, all three AI-based tools are oriented to linguistic precision; however, MS Copilot pays additional attention to improving sentences.

RQ2. Are there statistically significant differences among the three AI tools—ChatGPT, Gemini, and Microsoft Copilot—in the types of feedback provided?

One-way repeated measures ANOVA was used to find out whether there were any significant differences among the three AI technologies, namely ChatGPT, Gemini, and Microsoft Copilot, concerning the kinds of feedback provided in five categories of language skills: grammar, vocabulary, cohesion, sentence restructuring, and mechanics.

Table 3. One-way Repeated Measures ANOVA Results for AI Feedback Types

Feedback Type	df	F	p	η^2	Significant Difference ($p < .05$)
Grammar	(2, 68)	1.402	.253	.040	No
Vocabulary	(1.68, 57.19)	6.774	.004	.166	Yes
Cohesion	(1.31, 44.44)	2.710	.097	.074	No
Restructuring	(1.08, 36.79)	11.453	.001	.252	Yes
Mechanics	(1.53, 51.91)	11.351	< .001	.250	Yes

Table 3 summarizes the results of a one-way repeated-measures ANOVA performed to detect significant differences among the three AI tools concerning feedback types. As it is seen from the collected data, there were no significant differences between the three AI tools regarding grammar feedback, $F(2, 68) = 1.402$, $p = .253$, $\eta^2 = .040$, and cohesion feedback, $F(1.31, 44.44) = 2.710$, $p = .097$, $\eta^2 = .074$. Hence, the three tools provided students with similar amounts of feedback on grammar and cohesion.

At the same time, significant differences between the tools were revealed in vocabulary feedback, $F(1.68, 57.19) = 6.774$, $p = .004$, $\eta^2 = .166$; that is, there was a moderate effect of the AI tool on vocabulary-related feedback. Significant differences between the tools were also identified in terms of sentence restructuring, $F(1.08, 36.79) = 11.453$, $p = .001$, $\eta^2 = .252$, and mechanics feedback, $F(1.53, 51.91)$

=11.351, $p < .001$, $\eta^2 = .250$. Both restructuring and mechanics demonstrated large effect sizes, suggesting substantial variation among the AI tools in these feedback categories.

To summarize, the results obtained through using the three AI tools, it is safe to say that although all three tools showed an equal performance in terms of providing feedback on grammar and cohesion issues, they differed greatly regarding vocabulary development, sentence structure, and mechanics issues. It was especially noticeable in terms of sentence structure and mechanics.

Table 4. Post-hoc Pairwise Comparisons of Feedback Types Across AI Tools (Bonferroni Adjusted)

Feedback Type	Comparison	Mean Difference	Standard Error	df	t	pBonf
Vocabulary	ChatGPT – MS Copilot	-0.371	0.130	34	-2.85	.022
Vocabulary	Gemini – MS Copilot	-0.343	0.116	34	-2.96	.016
Restructuring	ChatGPT – MS Copilot	-1.143	0.328	34	-3.48	.004
Restructuring	Gemini – MS Copilot	-1.114	0.330	34	-3.38	.006
Mechanics	ChatGPT – Gemini	0.486	0.119	34	4.08	.001
Mechanics	ChatGPT – MS Copilot	0.286	0.113	34	2.53	.048
Mechanics	MS Copilot – Gemini	0.200	0.069	34	2.90	.019

Table 4 provides post-hoc pairwise comparisons of the feedback types which demonstrated significant differences in the repeated measures ANOVA. Based on the results, the significant differences observed from the analysis of variance were mainly due to comparisons of Microsoft Copilot in terms of vocabulary and sentence restructuring, and all pairwise comparisons of mechanics feedback.

Concerning vocabulary feedback, MS Copilot offered significantly more vocabulary feedback compared to ChatGPT (MDiff = -0.371, p = .022) and Gemini (MDiff = -0.343, p = .016). There was no significant difference between ChatGPT and Gemini. The results point out that MS Copilot is more concerned with vocabulary improvement compared to the two AI tools.

Regarding sentence restructuring, there were also some significant differences found between MS Copilot and other software. MS Copilot produced significantly more restructuring feedback than ChatGPT (MDiff = -1.143, p = .004) and Gemini (MDiff = -1.114, p = .006). It should be noted that no significant difference was found between ChatGPT and Gemini, which shows that MS Copilot was distinguished from them through its tendency to give more sentence-level advice.

In Addition, regarding mechanics feedback, all pairwise comparisons were significant. ChatGPT generated a significantly higher number of mechanics feedback than Gemini (MDiff = 0.486, p = .001) and MS Copilot (MDiff = 0.286, p = .048). Moreover, MS Copilot generated significantly more mechanics feedback than Gemini (MDiff = 0.200, p = .019). Consequently, a clear hierarchy of mechanics feedback is established in which ChatGPT has generated the highest number of feedback,

followed by MS Copilot and Gemini.

Overall, the post-hoc results show that MS Copilot is the main cause behind the differences found in feedback on vocabulary and sentence restructuring. It has also been found that ChatGPT is the best at providing feedback on mechanics. Thus, it is evident from the results that apart from giving feedback on a higher level, the three AI programs also vary in terms of what kind of feedback they give.

RQ3. How do these AI tools differ in terms of feedback style and depth?

To address the third research question, a thematic analysis was conducted to examine differences in feedback style and depth across ChatGPT, Google Gemini, and Microsoft Copilot. As shown in Table 5, the three AI tools exhibited distinct feedback patterns, suggesting different pedagogical orientations.

Table 5. Thematic Analysis of Feedback Generated by AI Tools

Category		ChatGPT	Gemini	MS Copilot
Feedback Style	Direct correction	Strong	Strong	Moderate
	Suggestion	Moderate	Strong	Moderate
	Rewriting	Limited	Limited	Very Strong
	List format	Frequent	Less frequent	Rare
	Paragraph feedback	Moderate	Strong	Strong
Depth of Feedback		Low	Medium	High
Main Function		Error correction	Teaching/guidance	Text improvement
Type of Tool		Corrective feedback tool	Instructional/teaching-oriented tool	Rewriting tool

From the table above, it can be seen that ChatGPT was primarily used as a tool for providing corrective feedback to learners. Indeed, the feedback provided by this tool was predominantly characterized by a focus on correction. The main emphasis was placed on grammar and mechanics, and most comments were list-based and aimed at the identification and correction of the errors committed by students. While the feedback could also include some suggestions and paragraph-level comments, the use of rewriting was quite limited in this case. For this reason, the feedback provided by ChatGPT was considered to be low in depth since it concentrated predominantly on error identification and correction.

Next, Gemini was characterized by a more instructional and educational approach. Similar to ChatGPT, it corrected many mistakes. In addition, it provided even more recommendations and remarks on a paragraph level. Unlike ChatGPT, this chatbot did not only correct mistakes but also explained issues to students and suggested how to improve their texts. Therefore, the feedback provided by Gemini was medium in depth.

In contrast, Microsoft Copilot operated quite differently, serving as an instrument of reworking rather

than providing feedback. Although the system provided corrections and recommendations from time to time, the major element of MS Copilot's operation was rephrasing the written sentences and even paragraphs. Therefore, the feedback could be defined as deep, because MS Copilot did not limit its operation to spotting mistakes but gave examples of how one can express certain thoughts. Such an approach can contribute to effective text improvement; however, at the same time, it limits students' opportunity to detect their mistakes on their own.

The results show that there are considerable distinctions between ChatGPT, Gemini, and Microsoft Copilot in terms of how these systems provide feedback. It appears that ChatGPT focuses mostly on identifying mistakes, while Gemini gives instructions and recommendations on writing. At the same time, MS Copilot rewrites the text rather than providing feedback. These differences suggest that each tool serves a distinct pedagogical function and may be suitable for different stages of writing development.

Discussion

The Findings revealed that ChatGPT, Gemini, and MS Copilot emphasized feedback on grammar, which indicates that writing systems based on artificial intelligence are especially effective in finding and fixing surface errors in language. ChatGPT and Gemini provided almost half of their feedback concerning grammar, and grammar was among the most frequent categories in feedback provided by MS Copilot. Lack of a statistically significant difference between the AI tools in providing grammar feedback indicates that grammatical correction has become a common feature of current writing assistants based on AI technology. These findings correspond to the results obtained by other researchers who have proved the efficiency of Automated Writing Evaluation and Generative Artificial Intelligence in terms of identifying issues connected with linguistic accuracy, grammar, and mechanics (McCarthy et al., 2022; Wei et al., 2023). Other recent review studies also stress that feedback provided by AI tools is more efficient in promoting form-focused writing skills, particularly those of EFL learners (Crosthwaite & Sun, 2026; Khatami & Suryati, 2026).

The other crucial finding from the current experiment was that the implementation of Microsoft Copilot as a sentence rephrasing tool was something distinctive for this particular application among ChatGPT and Gemini. MS Copilot generated much more feedback related to restructuring, and it exhibited clear tendencies towards rephrasing rather than editing. The above-mentioned conclusion coincides with the results of the recent studies that demonstrate how the functions of the generative AI systems are shifting towards text rewriting with increased readability and coherence (Alpar, 2025; Lin & Crosthwaite, 2024). On a similar path, Pack et al. (2025) found that large language models tend to use reformulations and modeled revisions while addressing student texts. The substantial discrepancy and large effect size in relation to restructuring feedback suggest that MS Copilot is a rewriting tool and not a tool of corrective feedback at all.

Differences in vocabulary and mechanics feedback were also found to be significant. MS Copilot provided significantly better feedback from the vocabulary perspective than ChatGPT and Gemini,

although ChatGPT gave the most mechanics feedback. The results coincide with previous research that confirmed the potential of generative AI technology in enhancing lexical richness through alternations and reformulations (Li et al., 2024; Teng & Huang, 2025). However, at the same time, it should be noted that ChatGPT performed better in giving mechanics feedback because ChatGPT paid attention mainly to the local editing errors (e.g., punctuation, spelling, and capitalization). This result is consistent with the conclusions of Al-Sofi (2026) who stated that AI systems excel at identifying sentence-level and mechanical errors, and Khatami and Suryati (2026), who pointed out that automatic feedback was the most effective in improving the surface-level writing accuracy.

The thematic analysis revealed that there were considerable differences among the three tools in terms of the style and depth of their feedback. ChatGPT was basically a corrective feedback tool that gave straightforward corrections. On the other hand, Gemini was an instructional feedback tool that not only provided corrections but also gave explanations and suggestions, thus making it conducive for learning. MS Copilot provided the most profound kind of feedback because of its extensive rewriting and reformulation at the paragraph level. Such outcomes are highly consistent with the findings of Guo and Wang (2024), who discovered that there is great variability in terms of style and educative value of AI feedback. In turn, Lin and Crosthwaite (2024) observed that there is a tendency for the generative AI to produce reformulation and explanation-based feedback rather than correction-based feedback.

In conclusion, it can be stated that the application of the AI technologies mentioned herein fulfills complementary pedagogical functions rather than competing ones. Specifically, ChatGPT is utilized for error detection and grammar verification, whereas Gemini works as a teaching aid tool and stimulates reflection and revisions. Lastly, Copilot is used as a rewriting tool. The results confirm the necessity of developing approaches to the human-in-the-loop integration of different AI technologies with consideration of pedagogical goals proposed by Guo and Wang (2024) and Pasha (2026). At the same time, recent literature reviews conducted by Khatami and Suryati (2026) and Lee et al. (2026) indicate that learner engagement, feedback literacy, and teacher mediation are crucial in the educational function of feedback provided with the help of AI technologies. Therefore, the optimal pedagogical strategy will involve the use of several AI tools together with teacher instructions.

Conclusion

The study offers an analysis of ChatGPT, Google Gemini, and Microsoft Copilot as feedback providers in EFL writing. The three tools consistently identified grammar, vocabulary, cohesion, sentence restructuring, and mechanics problems. However, there were no statistically significant differences between the three artificial intelligence systems as far as grammar and cohesion feedback was concerned, although the tools had different feedback priorities and pedagogical roles. In contrast, the most significant differences were in vocabulary and mechanics, and the most prominent differences were those concerning sentence restructuring. Overall, ChatGPT prioritized grammatical correctness and mechanics feedback, Gemini offered more balanced and elaborate feedback, while MS Copilot concentrated on

sentence restructuring and improvement. Such results are important to the growing body of literature on using AI technology in writing since they clearly illustrate that generative AI tools cannot be seen as pedagogically equal and serve different purposes in the process of writing. Rather than identifying a single "best" tool, the study highlights how different AI systems support different aspects of writing development.

Recommendations

Several important implications for language instructors and their students can be suggested based on the current study. First, the study implies that depending on the objectives of instruction, AI tools can be incorporated strategically in writing instruction, where ChatGPT can be helpful in error correction, Gemini is more likely to facilitate guided learning and revision, while MS Copilot will help develop textual fluency and sentence construction. Second, the findings of the research imply that teachers should incorporate various AI tools during instruction since using only one tool might prove less effective. Finally, it is evident from the study results that the existing trend concerning the necessity for using AI-generated feedback alongside the teacher's feedback is supported.

Limitations

There are a few limitations of the current study that should be mentioned. First, the study was performed on a relatively small sample of medical students at one institution. Therefore, its generalizability remains doubtful. Also, the features of the AI-generated feedback have been analyzed, but not the way learners used the feedback provided. There is a need for future studies to explore the effects of feedback generated by AI on students of various proficiency levels and writing genres. Also, it is suggested that other research can explore the effect of a combination of AI and teacher feedback on students' writing performance.

Ethics Statements

This study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki and complied with the ethical regulations and guidelines of Taif University. Ethical approval was obtained from the Ethics Committee of Taif University (Approval No. 47-419) before data collection. All participants were informed about the purpose and procedures of the study and provided written informed consent before participation. Participation was voluntary, and participants were informed of their right to withdraw from the study at any time without penalty. The confidentiality and anonymity of participant data were maintained throughout the research process, and all data were used exclusively for research purposes in accordance with institutional ethical guidelines and accepted standards for human-subject research.

Acknowledgements

The author would like to acknowledge Taif University students who participated in this study.

Generative AI Statement

The AI tool Microsoft Copilot is used for the purpose of proofreading, sentence rephrasing, and improving clarity. After using this AI tool, the researcher reviewed and verified the final version of the manuscript and took full responsibility for the content of the published work.

References

Alpar, Ö. (2025). Evaluating generative AI tools for improving English writing skills: A preliminary comparison of ChatGPT-4, Google Gemini, and Microsoft Copilot. *European Journal of Educational Research*, *14*(4), 1291-1308. https://doi.org/10.12973/eu-jer.14.4.1291

Al-Sofi, B. (2026). The effectiveness of AI-generated feedback in enhancing students' writing proficiency: A study of Google Gemini. *Technology in Language Teaching & Learning*, *8*, 103403. https://doi.org/10.29140/tltl.2026.103403

Bodaubekov, A., Agaidarova, S., Zhussipbek, T., Gaipov, D., & Balta, N. (2025). Leveraging AI to Enhance Writing Skills of Senior TFL Students in Kazakhstan: A Case Study Using" Write & Improve". *Contemporary Educational Technology*, *17*(1). https://doi.org/10.30935/cedtech/15687

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S.W., & Siemens, G. (2024). A meta-systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, *21*(4), 1-41. https://doi.org/10.1186/s41239-023-00436-z

Cheng, X., & Zhang, L. J. (2021). Teacher written feedback on English as a foreign language learners' writing: Examining native and nonnative English-speaking teachers' practices in feedback provision. *Frontiers in Psychology*, *12*, 629921. https://doi.org/10.3389/fpsyg.2021.629921

Crosthwaite, P., & Sun, S. (2026). Generative AI and L2 written feedback studies: A scoping review. *RELC Journal*, *57*(1), 207-219. https://doi.org/10.1177/00336882251386530

Farazouli, A., Cerratto-Pargman, T., Bolander-Laksov, K., & McGrath, C. (2024). Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices. *Assessment & Evaluation in Higher Education*, *49*(3), 363-375. https://doi.org/10.1080/02602938.2023.2241676

Gill, S. S., Xu, M., Patros, P., Wu, H., Kaur, R., Kaur, K., ... & Buyya, R. (2024). Transformative effects of ChatGPT on modern education: Emerging Era of AI Chatbots. *Internet of Things and Cyber-Physical Systems*, *4*, 19-23. https://doi.org/10.1016/j.iotcps.2023.06.002

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, *29*(7), 8435-8463. https://doi.org/10.1007/s10639-023-12146-0

Khatami, M. R., & Suryati, N. (2026). Automated writing evaluation and generative AI in Indonesian EFL Writing: A systematic review of effects, engagement, and pedagogical implications. *JEELS (Journal of English Education and Linguistics Studies)*, *13*(2), 633-660.

https://doi.org/10.30762/jeels.v13i2.7694

Kurtz, G., Amzalag, M., Shaked, N., Zaguri, Y., Kohen-Vacs, D., Gal, E., Zailer, G., & Barak-Medina, E. (2024). Strategies for Integrating Generative AI into Higher Education: Navigating Challenges and Leveraging Opportunities. *Education Sciences*, *14*(5), 503. https://doi.org/10.3390/educsci14050503

Labadze, L., Grigolia, M., & Machaidze, L. (2023). Role of AI chatbots in education: systematic literature review. *International journal of Educational Technology in Higher education*, *20*(1), 56, 1-17. https://doi.org/10.1186/s41239-023-00426-1

Lee, S., Choe, H., Zou, D., & Jeon, J. (2026). Generative AI (GenAI) in the language classroom: A systematic review. *Interactive Learning Environments*, *34*(1), 335-359. https://doi.org/10.1080/10494820.2025.2498537

Li, J., Huang, J., Wu, W., & Whipple, P. B. (2024). Evaluating the role of ChatGPT in enhancing EFL writing assessments in classroom settings: A preliminary investigation. *Humanities and Social Sciences Communications*, *11*(1), 1-9. https://doi.org/10.1057/s41599-024-03755-2

Lin, S., & Crosthwaite, P. (2024). The grass is not always greener: Teacher vs. GPT-assisted written corrective feedback. *System*, *127*. https://doi.org/10.1016/j.system.2024.103529

Link, S., Mehrzad, M., & Rahimi, M. (2022). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*, *35*(4), 605-634. https://doi.org/10.1080/09588221.2020

McCarthy, K. S., Roscoe, R. D., Allen, L. K., Likens, A. D., & McNamara, D. S. (2022). Automated writing evaluation: Does spelling and grammar feedback support high-quality writing and revision? *Assessing Writing*, *52*, 100608. https://doi.org/10.1016/j.asw.2022.100608

Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human-AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education*, *49*(5), 847-864. https://doi.org/10.1080/03075079.2024.2323593

Olsen, T., & Hunnes, J. (2024). Improving students' learning—the role of formative feedback: experiences from a crash course for business students in academic writing. *Assessment & Evaluation in Higher Education*, *49*(2), 129-141. https://doi.org/10.1080/02602938.2023.2187744

Pack, A., Hartshorn, K. J., Escalante, J., & Gillette, N. (2025). How well can GenAI (GPT-4) provide written corrective feedback on English-language learners' writing? *International Journal of English for Academic Purposes: Research and Practice*, *5*(1), 7-26. https://doi.org/10.3828/ijeap.2025.2

Pasha, M. K. (2026). Generative AI as an Automated Written Corrective Feedback Provider in EFL Academic Writing: A Mixed-Methods Investigation of Accuracy Gains, Feedback Quality, and Learner Engagement. *Journal of Artificial Intelligence and Computing*, *4*(1), 8-18. https://doi.org/10.57041/x2hg0114

Peláez-Sánchez, I. C., Velarde-Camaqui, D., & Glasserman-Morales, L. D. (2024). The impact of large language models on higher education: Exploring the connection between AI and Education 4.0.

Frontiers in Education, *9*. https://doi.org/10.3389/feduc.2024.1392091

Polakova, P., & Ivenz, P. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. *Cogent Education*, *11*(1). https://doi.org/10.1080/2331186X.2024.2410101

Shi, Y. (2021). Exploring learner engagement with multiple sources of feedback on L2 writing across genres. *Frontiers in Psychology*, *12*, 758867. https://doi.org/10.3389/fpsyg.2021.758867

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., ... & 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

Teng, M. F., & Huang, J. (2025). Incorporating ChatGPT for EFL writing and its effects on writing engagement. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*, *15*(1), 1-21.

Wang, Y. (2021). The Effectiveness and Issues of Using Feedback in Second Language Writing. *Review of Educational Theory*, *4*(4), 74-78.

Wei, P., Wang, X., & Dong, H. (2023). The impact of automated writing evaluation on second language writing skills of EFL learners. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1249991

Zhai, X. (2022). ChatGPT user experience: Implications for education. *SSRN Electronic Journal*. https://dx.doi.org/10.2139/ssrn.4312418

Zhao, X., Cox, A., & Cai, L. (2024). ChatGPT and the digitisation of writing. *Humanities and Social Sciences Communications*, *11*(1), 1-9. https://doi.org/10.1057/s41599-024-02904-x

Appendix 1

Feedback Analysis Framework

1- Feedback Type Categories

Code	**Type**	**Example**
G	Grammar	"internet waste → internet wastes"
V	Vocabulary	"videos not related → unrelated videos"
S	Sentence restructuring (rewriting a full sentence)	A smoother version: "Although the internet has downsides, such as addiction or distraction, its benefits far outweigh these risks when used responsibly."
C	Cohesion	"From other side → On the other hand"
M	Mechanics (spelling, punctuation, capitalization)	Spelling mistakes (e.g., oun → own, Forth → Fourth)

2- Feedback Style Categories

Feature	**Description**
Direct correction	Giving the correct form immediately
Suggestion	Offering alternative
Rewriting	Rewriting a full sentence
List format	Itemized corrections
Paragraph feedback	Continuous explanation

3- Depth of Feedback

Level	**Description**
Low	Only corrections
Medium	Corrections + some improvement
High	Corrections + explanations + rewriting