Article

Journal of Educational Computing Research
2025, Vol. 63(2) 431–463

DOI: 10.1177/07356331241307297
journals.sagepub.com/home/jec
Sage

The Effects of ChatGPT-Generated Feedback on Saudi EFL Learners' Writing Skills and Perception at the Tertiary Level: A Mixed-Methods Study

Abrar H. Alsofyani[1] and Amal M. Barzanji[2]

Abstract

Corrective feedback plays a critical role in enhancing writing skills among English as a Foreign Language (EFL) learners, but large class sizes often hinder the provision of personalized feedback. Generative AI tools such as ChatGPT have emerged as promising solutions that offer immediate and individualized feedback for writing. This study examined the effects of ChatGPT-generated feedback in comparison to teacher-generated feedback on Saudi EFL students' writing skills and perception at the university level. A mixed-method approach was employed to ensure comprehensive data collection and analyses. Using the pretest-posttest control group design, 102 participants were assigned to an experimental group, which received ChatGPT-generated feedback, and a control group, which received teacher-generated feedback. Pretests, posttests, and a perception survey were used to collect the data. ANCOVA, with pretest scores as a covariate, and content analysis of perception data were conducted. The results indicated no statistically significant differences in posttest scores between the experimental group and the control group, suggesting that ChatGPT-generated feedback was as effective as teacher-generated feedback. The perception survey

[1]English Language Institute, King Abdulaziz University, Jeddah, Saudi Arabia
[2]Applied College, Taibah University, Madinah, Saudi Arabia

Corresponding Author:
Amal M. Barzanji, Applied College, Taibah University, FGPR+CVM, Janadah Bin Umayyah Road, Tayba, Madinah 42353, Saudi Arabia.
Email: Ambarzanji@taibahu.edu.sa

revealed an overall positive attitude toward the use of ChatGPT for academic writing. The study provides empirical evidence supporting the effectiveness of generative AI tools in academic writing contexts.

Keywords
writing skills, corrective feedback, generative AI, English as a foreign language

Introduction

Written corrective feedback (WCF) plays a crucial role in writing classes, particularly in environments where English is a second or foreign language (L2/FL) (Bitchener & Storch, 2016). Although some scholars criticize the efficacy of feedback (Truscott, 1996), the majority still consider it beneficial for enhancing writing skills, even if the improvements are often limited to certain word-level linguistic features, such as articles (Bitchener & Ferris, 2012; Bitchener & Storch, 2016; Ferris, 1999). Numerous previous studies have demonstrated that WCF aids in L2 growth (Bitchener, 2008; Bitchener & Knoch, 2008, 2010a, 2010b; Ellis et al., 2008; Sheen, 2007; Sheen et al., 2009; Shintani & Ellis, 2013; Shintani et al., 2014). Specifically, Bitchener and Storch (2016) reviewed several studies on WCF and discovered that targeted WCF was more beneficial for students than untargeted WCF, especially when addressing specific forms of errors (Sheen et al., 2009). However, it should be noted that the treatment of untargeted errors was not systematic. In fact, no single approach can be considered supreme in correcting writing errors (Sheen et al., 2009).

Statement of the Problem

Despite all the benefits of WCF for developing students' writing skills, teachers often find it challenging to provide it for various reasons. For example, teachers often find WCF to be a laborious task, particularly at the tertiary level where the number of students is high. Giving each student individualized feedback takes a significant amount of time and effort, often overwhelming teachers. Additionally, the diversity in students' language proficiency levels necessitates tailoring feedback to individual needs, thereby increasing the complexity and time required. Teachers also face challenges in ensuring that students understand and effectively utilize the feedback given. Furthermore, balancing the provision of detailed feedback with other teaching responsibilities and administrative duties adds to the strain. There is also the issue of maintaining consistency and fairness in feedback, which can be difficult with large class sizes. However, with the advance of recent AI technologies such as chatbots and AI-enhanced writing software, teachers can facilitate and accelerate the WCF process and further provide customized feedback for each individual student.

Several studies (Ahmed, 2023; Algaraady & Mahyoob, 2023; Boudouaia et al., 2024; Guo & Wang, 2023; Han et al., 2023; Mahapatra, 2024; Masoudi, 2024) have suggested that ChatGPT is a promising feedback tool that can effectively improve ESL/EFL writing skills. These studies have presented empirical evidence on the effectiveness of ChatGPT in delivering individualized feedback on students' writing. Research on this topic has taken two primary directions: one focuses on assessing the quality of feedback provided by ChatGPT compared to teacher feedback (e.g., Algaraady & Mahyoob, 2023; Guo & Wang, 2023; Steiss et al., 2024), while the other examines the impact of these two types of feedback on enhancing the students' writing quality (e.g., Boudouaia et al., 2024; Mahapatra, 2024; Masoudi, 2024). For instance, a recent study by Steiss et al. (2024) compared the quality of ChatGPT's formative feedback with human feedback. They conducted the study on 200 secondary student essays, evaluating feedback on five aspects: criteria-based, clear directions for improvement, accuracy, prioritization of essential features, and supportive tone. The study utilized descriptive statistics and effect sizes to compare the quality of feedback between humans and AI. The study also examined differences based on essay quality and language status. The findings revealed that human raters provided higher quality feedback in most categories, except for criteria-based feedback, where ChatGPT performed slightly better. It is important to highlight that Steiss et al. concentrated on examining the differences in the quality of feedback offered by teachers compared to ChatGPT, rather than evaluating the improvements in students' writing resulting from the use of these distinct types of feedback. Additionally, following the other direction, Boudouaia et al. (2024) investigated the effectiveness of ChatGPT-4 in improving Algerian EFL students' writing skills at the tertiary level. To achieve this goal, pretest and posttest were administered to assess the improvements in the writing quality. The results showed a statistically significant improvement between pretests and posttests. The writing scores of the experimental group outperformed the control group. Furthermore, students' perceptions of using ChatGPT regarding its usefulness, ease of use, behavioral intentions, and attitudes were positive.

Rational of the Study

Although some studies (Boudouaia et al., 2024; Mahapatra, 2024; Masoudi, 2024) have examined how ChatGPT and teacher feedback contribute to improving students' writing, they did not directly compare teacher feedback with ChatGPT feedback or provide detailed insights into the parallel processes involved in both feedback types. Furthermore, while these studies assumed that the control group received traditional teacher feedback, they did not clarify the quantity or focus of this feedback. These studies primarily focused on the effectiveness of ChatGPT-generated feedback in improving students' writing quality. Furthermore, the students in these studies did not use a specific prompt to receive feedback; instead, they interacted with ChatGPT more freely. In contrast, the current study examined both teacher-provided and ChatGPT-generated feedback, ensuring that each type followed the same set of criteria.

Additionally, students in the current study used a detailed and specific prompt aligned with the writing criteria outlined in the rubric. The researchers used this rubric to evaluate writing quality in both the pretest and posttest.

Studies have also investigated the efficacy of teacher-provided feedback in comparison to ChatGPT-generated feedback (Algaraady & Mahyoob, 2023; Guo & Wang, 2023; Steiss et al., 2024). However, these studies primarily focused on analyzing the differences in the quality of feedback offered by teachers and ChatGPT, rather than on evaluating improvements in students' writing quality. Another distinguishing feature of the current study is the use of a structured survey tailored to the same writing criteria on which students received feedback in both teacher generated feedback versus ChatGPT generated feedback groups (TGF vs. CGF). This study specifically designed the survey items to align with the writing components assessed during the eight-week feedback period, in contrast to previous studies that largely examined students' perceptions of ChatGPT's overall acceptability, such as user satisfaction, ease of use, and interactive features (e.g., Ahmed, 2023; Boudouaia et al., 2024; Masoudi, 2024).

Literature Review

Theoretical Framework

To provide a strong foundation, this study is anchored in well-established theoretical frameworks. This study is based on two foundational theories: constructivism and sociocultural theory (specifically Vygotsky's Zone of Proximal Development). In combination, these theories underscore the pedagogical implications of utilizing both teacher and ChatGPT feedback to facilitate active, self-directed learning.

Constructivism

The constructivist framework emphasizes that learners actively construct knowledge through experience and reflection. According to constructivist principles (von Glasersfeld, 1996; Fosnot, 1996), learning is a dynamic, contextualized process where individuals build knowledge iteratively. In our study, both ChatGPT and teacher feedback enabled students to engage actively with their writing, revising their work based on provided feedback. This iterative process aligns with constructivist learning, where feedback is a critical tool for fostering autonomous learning. Learners engage in reflective and contextual learning by building upon unique experiences and teacher- or AI-guided insights, ultimately refining their writing skills and achieving greater independence in writing.

Sociocultural Theory

Rooted in the work of Vygotsky (Lantolf & Thorne, 2006), sociocultural theory (SCT) highlights the importance of social interactions and mediated learning within the Zone of Proximal Development (ZPD). This study applies SCT principles by examining how

teacher and ChatGPT feedback function as scaffolding tools, supporting learners within their ZPD. Through structured feedback sessions, students received guidance tailored to their evolving writing needs, facilitating progress beyond their initial capabilities. This approach exemplifies dynamic assessment within SCT, where both teacher and ChatGPT feedback provide continuous, need-based support that helps learners bridge the gap between current performance and potential skills.

In summary, these theoretical foundations illustrate the educational contributions of this study, which support a pedagogical model utilizing AI-generative feedback and provide insights into the effective use of AI as a scaffold for autonomous and contextual learning.

The Role of Automated Writing Software in Providing Written Corrective Feedback

Recent advancements in apps, software, and bots have made the process of providing corrective feedback (CF) easier than it was previously. Since the present study exclusively examined feedback generated by ChatGPT, the literature review will primarily focus on research utilizing AI-powered tools, including Automated Writing Evaluation (AWE) systems such as Project Essay Grade (PEG). For instance, Wilson and Czik (2016) conducted a quasi-experimental study to explore the effects of using Automated Writing Evaluation system, PEG, in conjunction with teacher feedback as opposed to using only teacher feedback through Google Docs. The study focused on the quantity, type, and level of teacher feedback, student writing motivation, and the students' quality of final-draft writing. The results indicated that there were no statistically significant differences in the quantity and type of feedback. However, the level of feedback showed statistically significant differences. For example, the findings revealed that the teacher gave a greater proportion of high-level feedback (e.g., sentence structure) in the combined condition using PEG + teacher feedback. In contrast, the teacher gave more lower-level feedback (e.g., formatting) in the Google Docs condition. In regard to the final draft writing quality, there were no statistically significant differences between groups. Wilson and Andrada (2016) conducted another study to determine if PEG could enhance writing quality for K-12 students in the United States, and the results showed that writing software can be particularly effective in improving writing skills.

In fact, various reviews have widely investigated the effectiveness of using AWE to improve writing skills (Fu et al., 2022; Stevenson & Phakiti, 2014; Zhai & Ma, 2023). For example, Stevenson and Phakiti (2014) conducted a critical review to analyze research on the effects of computer-generated feedback on the quality of students' writing. Their findings indicated that while there were some positive indications, the heterogeneity of studies, mixed results, and methodological issues limited the ability to draw firm conclusions about the effectiveness of AWE feedback. In a systematic review, Zhai and Ma (2023) also analyzed the effectiveness of AWE tools on the students' writing quality. The review included 26 primary studies conducted between 2010 and 2022. In contrast to Stevenson and Phakiti's (2014) review, this study found that there is a large and positive effect of AWE tools on the quality of students' writing.

Mohsen (2022) conducted a meta-analysis review to investigate the effectiveness of computer-generated written feedback on the development of L2 writing. The findings revealed a large effect of the non-AWE tools, which resulted in better improvements in writing quality compared to the AWE tools with a medium effect. Additionally, Fu et al. (2022) conducted a systematic review of 48 primary studies investigating the role of AWE in writing, focusing on the methodology, learners' types, feedback types and their applications, and learning outcomes. The findings revealed that although AWE feedback was helpful in improving writing skills, human feedback (both teacher and peer feedback) was still more effective.

The results of the review studies discussed in the previous section clearly show a slight discrepancy in the findings, which we can reasonably justify. Although Stevenson and Phakiti (2014) conducted a critical review of AWE, they noted several limitations in the body of research that prevented drawing firm conclusions about its effectiveness on writing quality. Specifically, they identified factors like the limited number of studies, participant diversity, research contexts, and varying methodologies as issues that constrained the ability to make definitive claims. AWE tools in 2014 relied on earlier artificial intelligence models, which were less sophisticated than those used in more recent tools.

However, recent meta-analyses (e.g., Mohsen, 2022; Zhai & Ma, 2023) indicate a larger effect of AWE on writing quality, likely due to advancements in AI technology and a more consistent research focus. Zhai and Ma (2023) found that university students, particularly L2 learners, benefited from AWE more than secondary students, with the greatest improvements seen in argumentative writing. Moreover, Mohsen (2022) observed that non-AWE tools without AI technology sometimes outperformed AWE tools, particularly for learners at beginner and intermediate proficiency levels, who showed notable progress, whereas advanced learners showed only slight improvement.

Fu et al.'s (2022) review revealed several crucial aspects and considerations related to the methodology, learner types, feedback types, their applications, and learning outcomes. Many limitations were discovered throughout the reviewed studies in Fu et al. article such as the increased focus on the quantitative methods, the limited number of participants, and the durations of the treatment sessions that did not exceed ten weeks. Overall, the review revealed a positive effect of the AWE on improving the writing quality, but this effect was found to be less significant compared to human feedback. As for the qualitative findings, although students showed positive attitudes towards using AWE, they were able to detect its flaws such as inaccurate, implicit, and general feedback. In summary, various factors may contribute to the inconsistencies observed in the reviewed literature on the recent AWE studies (Fu et al., 2022; Mohsen, 2022; Zhai & Ma, 2023) including the type of research methods, the number of participants, the feedback timing, types of writing, and learner proficiency.

ChatGPT and Its Benefits in Language Learning Classrooms

ChatGPT has received tremendous attention given its ability to provide quick and personalized feedback using AI technology at any time and for everyone. Therefore,

this section will start by defining ChatGPT. It will also discuss some of its uses in the language learning/teaching field. OpenAI (2024) states that GPT, or Generative Pre-trained Transformers, is based on generative models that employ deep learning technology. This technology manages a vast amount of data to train an AI system to perform any given task. The free version of ChatGPT is 3.5. A subscription is also available to access the advanced versions of ChatGPT (GPT-4 and GPT-4o). It is worth noting that ChatGPT can sync the user's history across his or her different devices. It also can recognize speech and enable voice input. It can provide instant answers, tailored advice, creative ideas, feedback, and learning opportunities (Yusuf et al., 2024). In fact, ChatGPT is considered a new kind of AI that utilizes Large Language Models (LLMs), and it is currently one of the most effective tools used in the field of language learning/teaching. This technology has been developed using an extensive number of human-generated texts. Consequently, it can predict human requests easily and then reply with human-like responses such as students' essays (Bonner et al., 2023).

Bonner et al. (2023) discussed some of the many services ChatGPT can offer to the language learning/teaching field. These services include summarizing language texts based on the learners' proficiency levels, editing grammatical and mechanical errors, creating narrative writing prompts, converting students' texts into presentation notes, providing lesson ideas, and creating texts for testing and reading practice that are suitable to different proficiency levels. In addition, according to Kim et al. (2023), ChatGPT is beneficial for language teachers as it can create different types of learning content in more than 12 languages that are suitable for different levels. For example, it can create lesson plans and evaluate students' writing. Additionally, the new version, ChatGPT-4o, can even teach (OpenAI, 2024). Both teachers and learners can benefit from ChatGPT services (Kim et al., 2023). Moreover, ChatGPT offers immediate responses, which provide a highly responsive and supportive learning atmosphere for language learners.

Despite the benefits of ChatGPT, it has raised academic concerns regarding integrity, such as potential biases, confidentiality issues, and risks of inaccurate information (Chan & Hu, 2023; Eke, 2023; Lim et al., 2023; Yusuf et al., 2024). ChatGPT's ability to generate accurate responses, including those involving technical skills like coding, can also lead to potential misuse by students and researchers (Lock, 2022); however, this is not always true. In fact, the accuracy and validity of AI-generated content are questionable, prompting calls for more research and the need to teach students proper use of AI tools in accordance with academic integrity principles (Peres et al., 2023).

Primary Studies on Using ChatGPT to Provide WCF for SL/FL Learners of English

In this section, the researchers focus on the primary studies investigating written corrective feedback (WCF) provided by ChatGPT to enhance the writing skills of second language (SL) and foreign language (FL) English learners. Among these studies, Algaraady and Mahyoob (2023) conducted a mixed-methods study to examine the efficacy of ChatGPT in detecting writing errors made by EFL students compared to

the feedback provided by instructors. The focus of the study was on analyzing the quality of the feedback itself, rather than assessing the improvements of the students' writing. Consequently, the study did not include feedback sessions or administer pretests and posttests. Instead, students' writing samples were collected and corrected by both teachers and ChatGPT. Researchers then analyzed these texts using four criteria: error identification, classification, analysis, and explanation. The analysis indicated that, while ChatGPT effectively identified surface-level errors—such as spelling, punctuation, grammar, and basic sentence structure issues—it struggled to detect deeper issues like coherence, argument structure, and content organization. The analysis demonstrated that human instructors provided feedback that was more accurate, particularly in addressing complex errors. Moreover, the findings revealed that ChatGPT's feedback cannot replace human feedback, especially when it comes to identifying pragmatic and deep structural issues.

In addition, Ahmed (2023) conducted a qualitative study to analyze Saudi learners' satisfaction with learning through feedback from both ChatGPT and their teachers. The study assessed four components of learning satisfaction: content, progress, ease of use, and interactive opportunities. On one hand, the results indicated that learners preferred teacher feedback over ChatGPT for interaction opportunities and ease of content learning. On the other hand, learners preferred ChatGPT due to its user-friendly nature compared to traditional teacher feedback methods. More importantly, the learners highlighted ChatGPT's limitations in facilitating learning progress without teacher guidance, emphasizing its reduced effectiveness when used independently.

Han et al. (2023) investigated the effectiveness of using ChatGPT as a feedback tool in EFL writing classes. The study focused on providing an interactive, guided learning process rather than a direct comparison between ChatGPT and human instructors. It emphasized the platform's ability to facilitate interaction rather than a side-by-side analysis. The findings suggested that students generally had a positive experience with ChatGPT, attributed to "its perceived helpfulness, trustworthiness, credibility, appropriateness of style/tone, performance, overall satisfaction, and referral intention." Moreover, Guo and Wang (2023) explored how ChatGPT supports EFL instructors by providing feedback in writing classes and examined the instructors' perceptions of using ChatGPT as a feedback tool. The study primarily analyzed ChatGPT's ability to provide effective feedback on EFL students' writing, highlighting its role as a supportive tool. Students' writing samples were collected, then evaluated by both teachers and ChatGPT. Similar to that study by Algaraady and Mahyoob (2023), this research did not include feedback sessions, pretests, or posttests. Instead, the focus was on analyzing the differences between feedback provided by ChatGPT and that offered by teachers. The findings indicated that ChatGPT consistently delivered extensive feedback across all evaluated criteria. In contrast, instructors primarily concentrated their feedback on content and language-related areas. The study emphasized differing approaches to feedback: ChatGPT tended to provide more directive feedback, while instructors offered more detailed insights, particularly in areas related to content and language.

The positive effect of ChatGPT in writing feedback extends into 2024, with studies reporting its continued influence. Steiss et al. (2024) investigated the effectiveness of ChatGPT in providing formative feedback on students' writing compared to feedback from human instructors. The findings indicated that the quality of feedback from human instructors was superior across all metrics, except for the criterion-based measure. This study concentrated on evaluating the quality of the feedback according to specific criteria. As a result, Steiss et al. did not incorporate pre- and post-tests or feedback sessions into their research design. The main focus was exclusively on the feedback itself rather than on the students' writing. Teachers initially assessed students' written work, and then the same texts were submitted to ChatGPT to generate feedback on the same components.

In contrast, Masoudi (2024), in a mixed-method study, investigated the impact of ChatGPT on enhancing specific aspects of English writing skills among Saudi university students. Masoudi focused on the improvements in the students' writing by administering a pretest and posttest to examine the differences. His analysis demonstrated a statistically significant improvement in the experimental group, particularly in areas such as grammar and vocabulary usage. This improvement was evidenced by the differences between pretest and posttest results. Furthermore, interviews and surveys revealed positive attitudes and perceptions toward the use of ChatGPT in enhancing writing proficiency. Notably, the primary goal of this study was to assess the effect of ChatGPT in improving the learners' writing proficiency rather than comparing the differences between ChatGPT and teachers' feedback.

Boudouaia et al. (2024) examined the efficacy of ChatGPT-4 in enhancing Algerian EFL students' argumentative writing skills at the university level. The researcher focused on evaluating the effect of the feedback by analyzing differences between the pretest and posttest results. The analysis revealed a statistically significant improvement between pretest and posttest. Importantly, the experimental group's writing on the posttest exceeded that of the control group. Additionally, students' perceptions of using ChatGPT were highly positive, particularly concerning its usefulness, ease of use, behavioral intentions, and attitudes. As in Masoudi's study (2024), Boudouaia et al. did not compare ChatGPT feedback with teacher feedback; instead, their primary focus was to assess the effectiveness of ChatGPT feedback in improving the students' writing.

Finally, Mahapatra (2024) conducted a study to explore the impact of using ChatGPT as a feedback tool on the writing abilities of ESL university students. The researcher administered a pretest, posttest, and delayed posttest to examine the differences in the students' writing. Similar to Masoudi's (2024) and Boudouaia et al.'s (2024) studies, this research did not directly compare teacher feedback with ChatGPT feedback. The findings indicated a statistically significant improvement in writing scores from the pretest to the posttest, with additional gains observed in the delayed posttest. These results suggest that incorporating ChatGPT feedback into writing instruction enhances students' writing abilities in terms of generating ideas, connecting sentences, and improving grammar. Additionally, discussion group feedback revealed that students appreciated receiving guidance from ChatGPT, particularly regarding

organization, content development, grammar, sentence structure, language accuracy, vocabulary choices, idea generation, peer collaboration, topic sentences, supporting details, main ideas, and conclusions. However, some students expressed concerns that relying on ChatGPT might reduce their motivation to think independently and potentially limit their creativity in content creation.

Based on the current literature, only a limited number of studies (e.g., Boudouaia et al., 2024; Mahapatra, 2024; Masoudi, 2024) have investigated the effectiveness of ChatGPT's feedback, primarily by evaluating improvements in students' writing. Although these studies utilized both experimental and control groups, their primary objective was to evaluate the effectiveness of ChatGPT in enhancing students' writing rather than comparing the impact of teacher feedback and ChatGPT-generated feedback on improving student's writing. Therefore, these studies did not provide a detailed description of the feedback given to the control group. Other studies (e.g., Algaraady & Mahyoob, 2023; Guo & Wang, 2023; Steiss et al., 2024) compared teacher and ChatGPT feedback; however, their primary focus was on evaluating the quality of the feedback rather than directly assessing improvements in writing. In contrast, the present study aimed to compare the impact of teacher and ChatGPT feedback on students' writing improvement through both pretest and posttest evaluations. To achieve this goal, the researchers employed a standardized and comprehensive rubric to assess four major components of writing: content, structure, grammar, and punctuation. The originality of the current study lies in its use of a tailored prompt, specifically designed around the same writing components included in the rubric.

In the current study, both groups (TGF vs. CGF) received feedback aligned with the same rubric, as detailed in the methodology section. By analyzing students' writing improvement, this study aims to contribute to existing research and determine whether ChatGPT feedback can outperform human feedback in effectiveness. The current study explored Saudi students' perceptions of utilizing ChatGPT feedback, specifically addressing detailed aspects of English writing, such as content, structure, grammar, and punctuation errors. Another distinctive feature of this study is the use of a structured and tailored survey focused on the same writing components on which the students received feedback. Most prior studies emphasized general acceptance features of ChatGPT, including ease of use, user satisfaction, and interactive opportunities (e.g., Ahmed, 2023; Boudouaia et al., 2024; Masoudi, 2024). In contrast, this study seeks to provide a more comprehensive understanding of ChatGPT's effectiveness as a feedback tool compared to teacher feedback. Table 1 below outlines the differences between this study and those cited in the literature review, categorized by objective, population, design, rubric, prompt, and results.

Methodology

Research Questions

This study was guided by the following research questions:

Table 1. Comparison of Primary Studies With the Current Study.

Studies	The current study	Algaraady and Mahyoob (2023)	Ahmed (2023)	Han et al. (2023)	Guo and Wang (2023)	Steiss et al. (2024)	Masoudi (2024)	Boudouaia et al. (2024)	Mahapatra (2024)
Objective	- Compared the effect of ChatGPT versus teacher feedback on students' writing improvement - Explored the students' perceptions regarding the use of ChatGPT as a feedback tool	Assessed ChatGPT's efficacy in detecting EFL writing errors versus instructors	Compared ChatGPT to human feedback, focusing on user satisfaction in teacher-mediated versus ChatGPT-assisted sessions	Explored ChatGPT's integration into EFL writing through a structured platform called RECIPE	- Analyzed ChatGPT's ability to support EFL instructors by providing feedback on writing. - Explored the perspectives of EFL teachers regarding ChatGPT feedback	- Evaluated the quality of ChatGPT's formative feedback compared to human instructors	- Investigated ChatGPT's impact on enhancing English writing skills - Explored the students' perceptions regarding the use of ChatGPT as a feedback tool	Investigated ChatGPT-4's efficacy in improving EFL students' argumentative writing and their perceptions of using ChatGPT	- Evaluated ChatGPT's efficacy as a formative feedback tool in improving writing in crowded classrooms. - Examined students' perceptions of ChatGPT's impact on their writing skills
Population	EFL university students (Saudi Arabia)	EFL university students (Saudi Arabia)	EFL university students (Saudi Arabia)	EFL undergraduate and graduate students (South Korea)	EFL Chinese instructors	Secondary school in the US (Native speakers and ELL)	EFL university students (Saudi Arabia)	EFL university students (Algeria)	ESL university students (India)
Design	Mixed methods	Mixed methods	Qualitative	Quantitative	Mixed methods	Quantitative	Mixed methods	Mixed methods	Mixed methods
Rubric	Assessed writing on content, structure, grammar, and punctuation	Error identification, classification, analysis, and explanation	Not specified	Not specified	Content, language, and overall feedback quality	Specific criteria, clarity, accuracy, essential features, and tone	Not specified	BI analytical evaluation scale, focusing on cohesion, coherence, grammar, and lexical accuracy	Content, organization, grammar, and vocabulary

(continued)

Table 1. (continued)

Studies	The current study	Algaraady and Mahyoob (2023)	Ahmed (2023)	Han et al. (2023)	Guo and Wang (2023)	Steiss et al. (2024)	Masoudi (2024)	Boudouaia et al. (2024)	Mahapatra (2024)
Prompt	Researchers designed tailored, detailed prompts for students aligned with the writing rubric	Used varied prompts to identify and analyze specific types of errors	Not specified	Not specified	Developed three prompts to generate and analyze feedback outputs	Focused on areas like supportive tone and clear directions in evaluating human versus ChatGPT feedback	Not specified	Not specified	Not specified
Results	- No statistically significant differences in posttest scores between the experimental group and the control group - Positive attitude toward the use of ChatGPT	- ChatGPT excelled in surface-level errors (e.g., spelling, grammar) - Human feedback was more accurate for deep-level errors (e.g., coherence, argument structure) - ChatGPT cannot replace human feedback, especially for complex issues	ChatGPT can support learning but cannot replace teachers without proper training, as it fails to enhance student progress due to insufficient learning satisfaction	Students had a positive experience with ChatGPT due to its helpfulness, trustworthiness, style, performance, and satisfaction	- ChatGPT provided extensive directive feedback - Instructors focused feedback on content and language with more detailed insights	- Human feedback was superior across most metrics except for criterion-based measures	- Significant improvement in experimental group (grammar, vocabulary) - Positive perceptions from students through surveys and interviews	Significant improvement in experimental group compared to control group - Positive perceptions regarding usefulness, ease of use, and attitudes	- Significant improvement in experimental group compared to control group - Positive perceptions regarding the use of ChatGPT

(1) What is the effect of ChatGPT generated feedback on Saudi EFL learners' writing skills?
(2) How do Saudi EFL learners perceive the effectiveness of ChatGPT feedback in improving their writing skills?

Research Design

This study employed an explanatory sequential mixed-methods design to explore the effectiveness of feedback provided by an AI-based tool, ChatGPT-3.5, as compared to conventional teacher feedback, on improving the writing proficiency of Saudi EFL learners. According to Creswell (2014), the study is classified as explanatory because it elaborates on the initial quantitative data results using qualitative data. The methodology is sequential, starting with the quantitative phase and subsequently transitioning to the qualitative phase (Creswell, 2014). The researchers initially utilized this model to conduct the quantitative portion of the study, and then refined the quantitative results using qualitative methods to gain deeper insights and a more comprehensive analysis. This methodological approach is known for its robustness in examining educational interventions from both quantitative and qualitative perspectives. It also offers a more comprehensive understanding of the phenomena under investigation (Creswell, 2008).

Participants were randomly assigned into two groups. The experimental group engaged with ChatGPT and received weekly feedback on their written assignments, while the control group received traditional teacher feedback on a weekly basis. Each student received up to 30 minutes of personalized feedback each week, tailored to their needs and guided by the rubric guidelines. The instructor provided this feedback through a combination of written and verbal guidance in one-on-one sessions. While feedback was personalized and rubric-guided, the traditional approach required substantial time and effort from the instructor, highlighting the value of this study in comparing AI-driven rapid feedback with conventional methods.

Context and Participants

The current study was conducted at a female campus of an English Language Institution in the western region of Saudi Arabia, with a group consisting of 102 Saudi female students at the tertiary level. The researchers used G*Power to calculate the recommended minimum sample sizes for the two samples at 0.05 alpha level, power of 0.8, and effect size of 0.6. The calculated minimum sample sizes for each sample were $n_1 = 45$ and $n_2 = 45$ (Total 90) and the researchers used $n_1 = 51$ and $n_2 = 51$ (Total 102). All the participants were native Arabic speakers. Their ages were 18 and 19 years old. All of the students had received a minimum of six years of English instruction prior to enrolling at the institution. The participants were drawn from three different classes but received instruction from one teacher with feedback provided during scheduled class times. The study lasted eight weeks, focusing on intermediate-level students who were learning to write a comprehensive paragraph. The participants' English language level was determined based on a standardized test administered by the English language institution.

Ethical Considerations

Ethical approval was secured from the institution's review board, with informed consent obtained from all participants. The anonymity and confidentiality of respondents were strictly maintained throughout the study. Prior to the experiment, the instructor explained the experimental objectives to the participants, ensuring that they understood the purpose and what was expected during their participation. Consent forms were distributed to gather explicit approval from all participants. To maintain confidentiality, each participant was given a pseudonym, and any personally identifying information was altered in the research documentation. The results were presented in aggregate to further reduce any risk of identifying individual participants.

Procedure

The experiment lasted approximately eight weeks, as shown in Table 2, and was conducted under the supervision of one teacher. The experimental group was then introduced to ChatGPT-3.5 in a structured training session, ensuring that the students were adequately prepared to use ChatGPT for their writing feedback. According to Mathur and Mahapatra (2022), it is recommended to train students before using any new technology in the classroom. The instructor provided the students with the appropriate ChatGPT prompt to receive the right feedback on each writing assignment. The feedback prompt was composed by the researchers following a standardized writing rubric used by the institution. The writing skills in the current study were evaluated based on four main criteria including structure (topic sentence, supporting details, concluding sentence), content (on topic, reasons and examples, linking words),

Table 2. Weekly Plan.

Weeks	Sessions	Feedback
1	Pretest for all participants the ChatGPT training session for the experimental group	
2	Writing session 1	ChatGPT (experimental group) individual teacher feedback (control group)
3	Writing Session2	Feedback continues for both groups
4	Writing session 3	Feedback continues for both groups
5	Writing session 4	Feedback continues for both groups
6	Writing session 5	Feedback continues for both groups
7	Writing session 6	Feedback continues for both groups
8	Posttest for both groups ChatGPT perception survey for the experimental group	

grammar (word order, personal pronouns, tenses, comparative and superlative), punctuation (full stop, capital letters, run-on, fragments, and spellings). Each component was scored on a standardized rubric from 1 to 5, resulting in a total writing score of 20 points. Table 3 presents each component and its subcomponents along with their corresponding scores. The feedback focused on helping students improve their writing skills according to their level.

Over the following six weeks, the writing sessions were conducted in which the experimental group ($n = 51$) received feedback from ChatGPT, while the control group ($n = 51$) received individual feedback from their teacher, focusing on the same writing

Table 3. Writing Rubric.

Criteria	Sub-criteria	Description	Score
Structure & content	Structure		5
	Topic sentence	Clear and relevant topic sentences introducing the paragraph's main idea	1
	Supporting details	At least 3 well-developed supporting details that are relevant and logically organized	3
	Concluding sentence	Clear concluding sentence that summarizes or reinforces the main idea	1
Content	Content		5
	On topic	Entire paragraph remains on topic with no irrelevant information	1
	Reasons, opinions, and examples	At least 3 reasons are given to support the main idea, with relevant examples or opinions	3
	Linking words	Uses at least 3 appropriate linking words to improve paragraph flow	1
Grammar & punctuation	Grammar		5
	Word order	Follows basic word order rules (subject-verb-object) with minimal errors	2
	Personal pronouns	Correct use of personal pronouns	1
	Tenses	Consistent and correct use of verb tenses	1
	Comparative and superlative	Accurate use of comparative and superlative structures (if used)	1
Grammar & punctuation	Punctuation		5
	Full stops	Full stops are used correctly at the end of sentences	1
	Capital letters	Correct use of capital letters at the beginning of sentences and for proper nouns	1
	Run-on sentences	Few to no run-on sentences	1
	Fragments	Avoids incomplete sentence fragments	1
	Spelling	Accurate spelling, especially for high-frequency words and words commonly used in academic settings	1
Total score			20

criteria as the ChatGPT group. In each session, the students in both groups had to write a paragraph about different topics structured according to the curriculum, receive feedback, and submit a revised version. By the end of week seven, all the students had already submitted six paragraphs with their revised copies. The experiment ended with a standardized post-test for both groups to measure any advancements in their writing skills. Finally, the experimental group received a perception survey regarding their opinions about the effectiveness of using ChatGPT as a feedback tool to improve their writing skills and their satisfaction and interaction with ChatGPT.

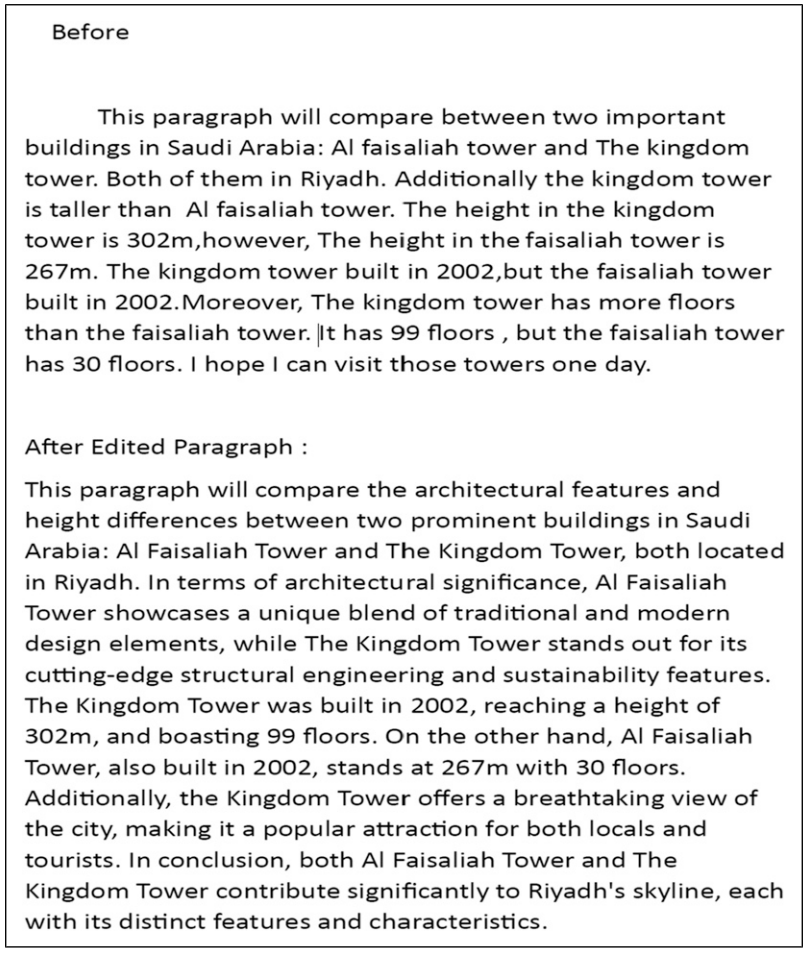
Tests

The present study administered two tests (pretest and posttest). Both tests were standardized and taken from the students' textbooks assigned by the institution. Therefore, the students were familiar with the test topics and writing genre. The tests adhered to a strict 70-min duration in an exam setting within the classroom, ensuring consistency and alignment with institutional assessment standards. The pretest was administered as a diagnostic assessment to determine the students' writing levels and to identify differences between the two groups. Students were required to write a descriptive paragraph. The pretest prompt was: "Describe the place where you live. Write about its positives and its negatives." The posttest was administered to assess any improvements in the writing of the two groups. The prompt required students to "Write a comparison of two buildings using the provided information." This prompt specifically instructed students to compare various aspects such as name, city, height, year, number of floors, number of lifts, cost, years to build, and any additional information about the buildings.

Creating ChatGPT Prompt

The creation of the GPT prompt went through several attempts and adjustments until the researchers agreed on the final prompt. Then, the final prompt was evaluated by an expert in the ESL field. The writing criteria used in the prompt were taken from the English institution writing rubric. Figure 1 below shows an example of a student paragraph before and after ChatGPT feedback. Below is the final prompt used in the study.

> As an English language instructor, I would like you to edit the following paragraph and make sure that it has a clear structure and correct grammar. My level is A2 according to the CEFR scale. Edit only if there are errors. Please, find the structure errors such as an unclear/missing topic sentence, lack of supporting details, an unclear or missing concluding sentence. Highlight all the errors and then edit them in a separate paragraph. The paragraph should include at least 3 supporting sentences that support the main idea. These details can include examples, evidence, or explanations that clarify my point. Make sure I


Before

This paragraph will compare between two important buildings in Saudi Arabia: Al faisaliah tower and The kingdom tower. Both of them in Riyadh. Additionally the kingdom tower is taller than Al faisaliah tower. The height in the kingdom tower is 302m,however, The height in the faisaliah tower is 267m. The kingdom tower built in 2002,but the faisaliah tower built in 2002.Moreover, The kingdom tower has more floors than the faisaliah tower. It has 99 floors , but the faisaliah tower has 30 floors. I hope I can visit those towers one day.

After Edited Paragraph :

This paragraph will compare the architectural features and height differences between two prominent buildings in Saudi Arabia: Al Faisaliah Tower and The Kingdom Tower, both located in Riyadh. In terms of architectural significance, Al Faisaliah Tower showcases a unique blend of traditional and modern design elements, while The Kingdom Tower stands out for its cutting-edge structural engineering and sustainability features. The Kingdom Tower was built in 2002, reaching a height of 302m, and boasting 99 floors. On the other hand, Al Faisaliah Tower, also built in 2002, stands at 267m with 30 floors. Additionally, the Kingdom Tower offers a breathtaking view of the city, making it a popular attraction for both locals and tourists. In conclusion, both Al Faisaliah Tower and The Kingdom Tower contribute significantly to Riyadh's skyline, each with its distinct features and characteristics.

Figure 1. Example of a student paragraph.

used a transitional word to end my paragraph. Put the structure errors in a table, explain the errors, give examples from my own writing, and provide corrections.

For the grammatical errors, focus on these error categories: Capital letters, full stops, spelling, verb tense, basic word order (subject-verb-object), run-on sentences, fragments, and personal pronouns. Edit only if there are errors. Put the grammatical errors in a separate table and categorize them, explain each error, and give examples for the corrections.

This is my paragraph:

Instruments and Reliability. The researchers utilized two primary instruments. Initially, to address the first research question — 'What is the effect of ChatGPT generated feedback on Saudi EFL learners' writing skills?'—the researchers conducted standardized pretest and posttest. For the second research question— ' How do Saudi EFL learners perceive the effectiveness of ChatGPT feedback in improving their writing skills? ' — the researchers employed a perception survey specifically designed by the researchers to address the unique aspects of writing feedback as applied in this study (see Appendix A). Subsequently, the survey was reviewed and validated by three EFL experts and modified based on their feedback. It was also translated into Arabic to avoid language barriers. The survey was distributed after the experiment was done to gather the students' subjective perceptions of ChatGPT's role in improving writing skills. The survey items used a Likert scale ranging from 1 to 5, where 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree. The survey was then digitized using Google Forms to facilitate data collection. To assess the reliability of the survey results, the researchers calculated Cronbach's Alpha, which was 0.86, indicating a very good reliability (Tavakol & Dennick, 2011).

Data Analysis

Two raters, who are experts in the EFL writing field, graded the pretest and posttest. Each student's maximum score was 20, based on the institution's writing rubric. The researchers calculated the mean of the scores assigned by the raters and ensured consistency through an Interclass Correlation Coefficient (ICC) analysis. All ICC values exceeded 0.9, indicating strong inter-rater agreement (ChatGPT Pretest: 0.993; Teacher Feedback Pretest: 0.988; ChatGPT Posttest: 0.984; Teacher Feedback Posttest: 0.943).

The researchers then analyzed the scores, setting the alpha level at 0.05 for all statistical tests. The analysis began with assumptions tested for normality and homogeneity. The Shapiro-Wilk Normality Test confirmed the data's normal distribution, while Levene's Test for Equality of Variances verified homogeneity.

To address the first research question, the researchers conducted an independent samples t-test to detect statistically significant differences in pretest scores between the two groups. Since a significant difference was found, an ANCOVA test was performed using pretest scores as a covariate. The researchers also calculated the effect size of the posttest and assessed statistical power using the G*Power program. For the second research question, the researchers employed both quantitative and qualitative analytical approaches. For the quantitative analysis, a descriptive analysis was conducted to calculate the weighted scores from the survey responses, and Cronbach's Alpha was used to assess the internal consistency of the survey items. For the qualitative approach, content analysis was applied to identify common themes and insights from students' feedback on the use of ChatGPT. This approach provided a comprehensive overview of student feedback, well-aligned with the study's exploratory aims.

Results

RQ1- what is the Effect of ChatGPT Generated Feedback on Saudi EFL Learners' Writing Skills?

For a preliminary analysis, the researchers employed the Shapiro-Wilk test to evaluate whether the data followed a normal distribution and Levene's test to assess the equality of variances. The Shapiro-Wilk results (CGF = 0.859; TGF = 0.819) demonstrated that the data were reasonably normal. However, Levene's test revealed a statistically significant p-value ($p < .0001$), indicating significantly different variances.

The researchers then examined whether there was a statistically significant difference between the pretest scores for the CGF and TGF groups using an independent samples t-test. The results (Table 4) showed that the TGF group scored higher (M = 13.27, SD = 1.70) than the CGF group (M = 10.83, SD = 3.00). The variances were unequal; thus, a t-test was performed without pooling variances, confirming a statistically significant difference ($p < .0001$).

After receiving feedback, both groups improved their posttest scores. The CGF group's mean increased to 16.39 (SD = 2.42), while the TGF group improved to 17.55 (SD = 1.01) as shown in Table 5.

An analysis of the difference in posttest means (Table 6) revealed a statistically significant difference ($\mu 1 - \mu 2 = 2.45$, $p < .0001$). However, the researchers employed ANCOVA to account for the significant differences in pretest scores between groups.

After accounting for pretest scores as a covariate, the ANCOVA revealed no statistically significant difference in posttest scores between the CGF and TGF groups (p = .052, Table 7). This result addressed initial disparities in pretest scores, ensuring a fair comparison of treatment effects.

RQ2- How Do Saudi EFL Learners Perceive the Effectiveness of ChatGPT in Improving Their Writing Skills?

The researchers examined students' perceptions regarding the effectiveness of using ChatGPT feedback to improve their writing skills. Research question 2 was analyzed using both quantitative and qualitative methods.

Survey Results. As summarized in Table 8, the students reported a high level of satisfaction with ChatGPT's ability to improve various aspects of their writing skills. The students reported an average mean score of $M = 4.5$ for how ChatGPT helped with topic sentences. The mean score for supporting sentences was $M = 4.0$, and for concluding sentences, it was $M = 4.4$. Grammar received an average mean score of $M = 4.4$. Correct verb tenses and correct word order had mean scores of $M = 4.3$ and $M = 4.2$, respectively. Similarly, correcting personal pronouns received $M = 4.2$, and avoiding spelling errors was rated $M = 4.4$. Correcting fragments and run-on sentences had a mean score of $M = 4.3$, while improving the use of capitalization and full stops scored

Table 4. Preliminary Analysis of the Pretest Scores.

Test	*n*	*M*	Variance	*SD*
ChatGPT pretest	51	10.83	8.99	3.00
Teacher feedback pretest	51	13.27	2.88	1.70

Table 5. Descriptives Results of the Posttest Scores.

Treatment	*N*	*M*	*SD*	*SE*	Coefficient of variation
ChatGPT	51	16.39	2.42	0.338	0.147
Teacher feedback	51	17.55	1.01	0.141	0.057

Table 6. The Difference Between the Means of the Two Groups.

Difference	*N*	*SE*	*DF*	*T*-stat	*p*-value
μ1−μ2	2.45	0.48	79.1	5.08	<0.0001

Table 7. ANCOVA Results of the Posttest Scores.

Cases	Sum of squares	df	Mean square	*F*	*p*
Treatment	12.916	1	12.916	3.874	0.052
Pretest	12.706	1	12.706	3.811	0.054
Residuals	330.079	99	3.334		

Note. Type III Sum of Squares.

$M = 4.0$ and $M = 4.3$, respectively. The students reported an average mean score of $M = 4.5$ for the overall improvement of their writing skills, and the ease of use of ChatGPT scored the highest average mean of $M = 4.6$. The experience of receiving feedback was rated $M = 4.4$, and the recommendation to use ChatGPT for writing correction had a mean score of $M = 4.5$. However, the preference for ChatGPT over human instructors was lower, with a mean score of $M = 3.4$. The perceived accuracy of ChatGPT's feedback received a mixed response, with an average mean score of $M = 3.8$. Awareness of ethical issues such as plagiarism was rated $M = 4.2$. The overall reliability of the survey responses was calculated with a Cronbach's Alpha of 0.86, indicating high internal consistency and reliability in the responses

Open-Ended Questions Results. The participants discussed their perceptions of using ChatGPT to develop writing skills, covering areas such as improvement in various writing abilities and the advantages and disadvantages of using ChatGPT, as

Table 8. Survey Questions With the Results, Starting From the Highest to the Lowest Mean Weighted Score.

Rank	Question	Weighted score
1	Easy to use	4.6
2	Topic sentence	4.5
3	I recommend ChatGPT for writing correction	4.5
4	Improve overall writing skills	4.5
5	Concluding sentence	4.4
6	Grammar	4.4
7	Avoid spelling errors	4.4
8	Positive experience receiving feedback	4.4
9	Correct verb tenses	4.3
10	Correct fragments and run-on sentences	4.3
11	Improve use of full stops	4.3
12	Correct word order	4.2
13	Correct personal pronouns	4.2
14	Aware of ethical issues such as plagiarism	4.2
15	Improve use of capitalization	4.0
16	Supporting sentences	4.0
17	I think ChatGPT feedback accurate	3.8
18	Prefer ChatGPT over my instructor	3.4

demonstrated in Table 9. The students found the ChatGPT experience both amazing and enjoyable. They believed that ChatGPT fostered their independence as college students, enabling them to learn autonomously through its valuable feedback. Additionally, students found that all the feedback was useful for learning skills, starting from topic sentences, details, and concluding sentences to grammar and spelling. ChatGPT provided immediate and fast feedback, which they appreciated being available at any time. ChatGPT assisted them in identifying all errors and providing explanations and justifications for each piece of feedback, all with just the press of a button. They recommended it to all students who face difficulties with writing.

Moreover, the students appreciated how ChatGPT corrected their punctuation errors, such as full stops, capital letters, and commas. They noted that ChatGPT effectively organized their paragraphs by logically connecting supporting details. This helped them learn to construct specific supporting details and consider alternative, improved sentences. They also valued how ChatGPT's editing of tenses drew their attention to the appropriate tense usage for their context. They recommended ChatGPT to all students who struggle with writing, as it also proved helpful in other courses. Students also found that ChatGPT was a useful tool for reluctant learners. Some students felt uncomfortable sitting with their teacher and showing their mistakes. However, they felt relaxed when writing and receiving feedback from ChatGPT. This encouraged them to write more without the fear of making mistakes.

Table 9. Summary of the Results of the Open-Ended Questions.

Aspect	Details
Advantages	Participants found ChatGPT to be an amazing and enjoyable experience
	Encourages independent learning, making students feel like self-reliant college learners
	Provides useful feedback on various writing skills, including topic sentences, supporting details, concluding sentences, grammar, and spelling
	Provides immediate and fast feedback anytime
	Identifies errors, provides explanations, and justifies corrections with a single button press
	Improves punctuation accuracy, including full stops, capital letters, and commas
	Helps organize paragraphs with connected supporting details and improves sentence construction
	Corrects tenses consistently, aiding students in learning proper tense usage
	Beneficial for writing assignments in other courses as well
	Serves as a great tool for reluctant students who feel uncomfortable sharing their mistakes with teachers
Disadvantages	Corrections can be unrealistic
	Adds extra sentences not in the original text
	Some students disagree with corrections
	Sometimes rewrites sentences in a fancy way, adding unneeded information
	Some students feel little improvement in writing
	Alters the original meaning at times

Regarding the disadvantages, students observed that some corrections made by ChatGPT were unrealistic. For instance, it occasionally introduced sentences that were not part of the original text. Participants noted that ChatGPT sometimes altered the meaning of a paragraph by adding unnecessary details. Some students disagreed with its corrections, arguing that even when a sentence was initially correct, ChatGPT would attempt to rewrite it more elaborately, often introducing extraneous information. One student remarked, '*I don't advise anyone to completely depend on ChatGPT. Yes, it may provide excellent examples and well-worded paragraphs, but it can make mistakes and isn't always accurate in academic writing.*' A few students also felt that their writing had not improved significantly.

Discussion

RQ1- what is the Effect of ChatGPT Generated Feedback on Saudi EFL Learners' Writing Skills?

The results of the ANCOVA test revealed no statistically significant difference between the two groups (CGF vs. TGF) regarding the increase in posttest scores. This indicates that ChatGPT-generated feedback was as effective as the teacher-generated feedback provided in this study. It is unsurprising that receiving targeted feedback from teachers is

likely to enhance students' writing skills (Bitchener & Storch, 2016; Sheen et al., 2009). According to Ferris (2011), error feedback provided by teachers can indeed help students produce accurate texts. This type of feedback has proven to be more helpful than generic feedback applicable to all students, as highlighted by Ferris (2008). However, providing individualized feedback in large classes presents a significant challenge, as correcting students' errors can be both exhausting and time-consuming, potentially leading to teacher burnout (Ferris, 2011). Numerous researchers have questioned the adequacy of traditional teacher feedback, describing it as "incomplete, idiosyncratic, erratic, and inaccurate" (Ferris, 2011, p. 20). In contrast, technological tools and apps such as ChatGPT do not experience weariness and can be highly effective in providing feedback. The findings of the current study indicate that ChatGPT can be quite effective in delivering feedback that, in some cases, matches the quality of feedback provided by teachers and leads to improvements in students' writing quality. These findings align with the results of Boudouaia et al. (2024), Mahapatra (2024), and Masoudi (2024), who also recognized ChatGPT's effectiveness in enhancing learners' writing quality. ChatGPT offers immediate feedback and individualized assistance (Han et al., 2023; Xiao & Zhi, 2023), benefiting both diverse language learners and overburdened teachers. This is particularly beneficial for university teachers handling large classes.

In fact, individualized assistance aligns with the constructivist theory, as it empowers students to become autonomous and active learners who iteratively build and construct their knowledge. The integration of AI-based feedback tools like ChatGPT allows students to receive unlimited WCF at their convenience. ChatGPT feedback plays a significant role in scaffolding and supporting learners within their ZPD. It provides guidance tailored to their specific writing needs and thereby facilitates their progress. This support aligns with SCT theory and the concept of ZPD, reinforcing its educational value. ChatGPT has proven its ability as a competent teaching assistant, helping educators by providing feedback and reducing their workload. Therefore, we highly recommend that writing teachers integrate ChatGPT into their classrooms, while also ensuring they receive professional development training to effectively utilize this technology.

Educators need to be aware of ChatGPT's flaws. First, ChatGPT frequently fails to provide accurate corrections, as highlighted by studies such as those by Guo and Wang (2023), Fu et al. (2022), Steiss et al. (2024), and Xiao and Zhi (2023). In the current study, the researchers observed a significant portion of inaccurate feedback from ChatGPT, as illustrated in Table 10. For instance, ChatGPT incorrectly changed the preposition in the sentence "the height in the Kingdom Tower is 203m" from "in" to "off," which is not a word order error. Furthermore, the feedback from ChatGPT suggested that the student's paragraph lacked supporting details, yet the original text did include such details. ChatGPT also mistakenly classified the sentence "Additionally, the Kingdom Tower is taller than Al Faisaliah Tower" as a fragment error. This error identification is incorrect; the sentence simply requires a comma after the introductory phrase. Obviously, ChatGPT struggles to accurately identify the type of error. In the study by Xiao and Zhi (2023), students were hesitant to accept ChatGPT feedback without questioning its accuracy. They critiqued the limitations and usability of ChatGPT feedback across various learning

aspects. Similarly, Kim et al. (2023) noted that ChatGPT has limitations in assessing students' writing. Guo and Wang (2023) observed that ChatGPT provided off-task feedback in their study, which did not align with established feedback categorizations. ChatGPT also provided irrelevant comments and failed to identify certain writing errors. These findings align with the current study's results, highlighting the need for caution and critical evaluation when using ChatGPT for educational purposes.

Second, ChatGPT tends to confuse learners, particularly those with lower and intermediate proficiency levels. Sometimes, students with lower language proficiency struggle to comprehend ChatGPT feedback, as reported by Fu et al. (2022). The extensive and occasionally irrelevant feedback provided by ChatGPT may contribute to this confusion. In the current study, the researchers observed that ChatGPT feedback was lengthy, complicating the students' ability to understand the corrections. As shown in Table 10, ChatGPT inaccurately criticized the clarity of the topic sentence and the presence of supporting sentences; however, the researchers found that the students had written an adequate topic and supporting sentences. It would have been sufficient if ChatGPT had merely commented on the concluding sentence. Similarly, Guo and Wang (2023) highlighted that the lengthy feedback from ChatGPT was particularly challenging for students with low proficiency.

To conclude, while ChatGPT feedback cannot replace teacher feedback, its effectiveness in enhancing learners' writing quality is well supported by research from Boudouaia et al. (2024), Mahapatra (2024), and Masoudi (2024). The current study has demonstrated that ChatGPT, despite its limitations, improves students' writing scores. ChatGPT feedback was as effective as teacher feedback. When used appropriately, ChatGPT can serve as a valuable resource for both teachers and students. However, due to its inherent limitations and possible inaccuracies, students should not rely entirely on ChatGPT for feedback. Teachers should encourage students to critically evaluate and thoroughly analyze ChatGPT's responses.

RQ2- How Do Saudi EFL Learners Perceive the Effectiveness of ChatGPT in Improving Their Writing Skills?

Based on the survey results, the ease of ChatGPT use received the highest weighted mean score. This finding aligns with the studies by Ahmed (2023) and Boudouaia et al. (2024), where participants also rated the ease of use highly. This can be attributed to the user-friendly design of ChatGPT, which facilitates easy interaction for students. In the open-ended section, the students expressed appreciation for the rapid and immediate feedback from ChatGPT, highlighting the convenience of receiving answers at the press of a button. According to OpenAI (2024), ChatGPT can provide immediate answers and tailor feedback to users' needs. It's easy-to-navigate interface ensures a smooth and pleasant learning experience. In fact, the learners in the current study described their experience using ChatGPT as enjoyable and amazing.

The second highest scores were given to improving topic sentences, ChatGPT's ability to correct students' errors, and enhancing overall writing skills. The students believed that among all the structural errors, ChatGPT was particularly helpful in guiding them to write

Table 10. ChatGPT Detailed Corrections and Potential Errors.

Error category	Explanation of error	Example from writing	Correction example
Capital letters	Incorrect use of capitalization	"Hight" instead of "height."	"The height of the Kingdom tower is 302m, whereas the height of Al Faisaliah tower is 267m."
Full stops	Missing or incorrect use of full stops	Lack of a full stop after "Riyadh."	"Both of them are in Riyadh."
Verb tense	Inconsistency in verb tenses	"Built" instead of "build" and inconsistency in verb tenses throughout the paragraph	"The Kingdom tower was built in 2002, while the Faisaliah tower was built in the same year."
Basic word order	Incorrect word order in sentences	"However, the height in the Kingdom Tower is 302m" should be "However, the height of the Kingdom tower is 302m."	"However, the height of The Kingdom Tower is 302m."
Personal pronouns	Incorrect use of personal pronouns	"The hight in the Kingdom Tower is 302m" instead of "the height of The Kingdom Tower is 302m."	"The height of The Kingdom Tower is 302m."
Fragments	Incomplete sentences	"Additionally The Kingdom Tower is taller than Al Faisaliah tower."	"Additionally, the Kingdom tower is taller than Al Faisaliah tower."
Run-on sentences	Incorrect use of run-on sentences	"It has 99 floors, but the Faisaliah tower has 30 floors."	"It has 99 floors. In contrast, the Faisaliah tower has 30 floors."
Unclear/ Missing topic sentence	The paragraph lacks a clear topic sentence that introduces the main idea	Absence of a topic sentence	"This paragraph will compare the architectural features and height differences between two prominent buildings in Saudi Arabia: Al Faisaliah tower and the Kingdom tower, both located in Riyadh."

(continued)

Table 10. (continued)

Error category	Explanation of error	Example from writing	Correction example
Lack of supporting details	The paragraph lacks supporting details such as examples, evidence, or explanations to clarify the main points	Absence of supporting details	"In terms of architectural significance, Al Faisaliah tower showcases a unique blend of traditional and modern design elements, while the Kingdom tower stands out for its cutting-edge structural engineering and sustainability features."
Unclear/ Missing concluding sentence	The paragraph does not have a clear concluding sentence that summarizes the main points discussed	Absence of a concluding sentence	"In conclusion, both Al Faisaliah tower and the Kingdom tower contribute significantly to Riyadh's skyline, each with its distinct features and characteristics

better topic sentences. This result is supported by Mahapatra's (2024) study, where students reported that ChatGPT helped them improve the structure of their writing, including writing a good topic sentence. Moreover, the students believed that ChatGPT enhanced their overall writing skills and, therefore, recommended it to other students seeking assistance with writing corrections. These results align with the findings of Boudouaia et al. (2024), Mahapatra (2024), and Masoudi (2024), which also indicated students' positive perceptions regarding the use of ChatGPT for writing feedback. The results demonstrated that ChatGPT effectively helps students improve their writing, suggesting that the latest generative AI applications have evolved to a level where they can compete with human capabilities in certain instructional contexts.

The least weighted scores were given to the accuracy of ChatGPT in providing feedback and preferring ChatGPT over teacher feedback, respectively. This can be attributed to the inaccurate responses provided by ChatGPT, as discussed in the previous section. Numerous researchers, including Guo and Wang (2023), Fu et al. (2022), Steiss et al. (2024), and Xiao and Zhi (2023), have noted that ChatGPT often fails to provide accurate feedback. This issue might be explained by the phenomenon of AI hallucinations. According to IBM (2024), AI hallucinations occur when a chatbot perceives non-existent patterns and consequently generates responses that are inaccurate or nonsensical. Users usually expect accurate output to their prompts; however, AI algorithms often provide inaccurate outputs due to bias/inaccuracy, training data, overfitting, and high model complexity. Similar to humans and animals, AI can hallucinate and see figures that do not exist like when humans see faces in the clouds or the moon. Interestingly, the statement

"preferring ChatGPT over my instructor" received the lowest weighted score. This finding aligns with Ahmed's (2023) study, where participants preferred receiving feedback from their teachers rather than from ChatGPT. This preference may be attributed to the ease of interaction with human instructors. Despite the advancements in AI feedback tools, interacting with machines can still be perceived as tedious. Finally, in the current study, students generally value their teachers' feedback more (Ferris, 2011) and often disregard ChatGPT's feedback, viewing it as inaccurate (Xiao & Zhi, 2023).

Recommendations and Future Research Directions

This study assessed the effectiveness of ChatGPT feedback compared to traditional teacher feedback in enhancing the writing skills of Saudi female EFL students at the university level. The researchers have several recommendations to offer for future research directions. This study used a customized prompt based on learners' proficiency levels, which shaped ChatGPT's feedback. Therefore, while ChatGPT provided useful feedback in this context, these findings are specific to the conditions of this study. Researchers should view ChatGPT's effectiveness as a feedback tool as dependent on its use and context, not as an absolute characteristic of the tool. Future research could explore how different prompt designs and instructional contexts impact the effectiveness of ChatGPT feedback.

The participants in the current study were at an intermediate proficiency level. Replicating this study with students of advanced proficiency and providing opportunities for more in-depth interaction with ChatGPT may produce different results. Future studies could incorporate a delayed posttest to measure the long-term retention of skills after the intervention. Future researchers could also explore the specific characteristics of ChatGPT feedback in comparison to teacher-provided feedback, with a focus on how aspects such as feedback length, complexity, and clarity influence students' ability to understand and apply corrections. Furthermore, the G*Power analysis deemed the study sample size sufficient, but a larger sample could potentially alter the outcomes. The current study lasted for eight weeks, aligning with the university's academic semester. Although this duration offered insights within a typical semester, extending future studies to 12 weeks or more could allow for a more comprehensive assessment of prolonged AI interaction's learning outcomes.

Finally, the researchers recommend comparing the effectiveness of different AI feedback tools, not just ChatGPT, to determine which features are most beneficial for language learning. This could aid in developing a more tailored AI tool that addresses specific educational needs. In the current study, the researchers used ChatGPT version 3.5; future studies might consider using the latest versions, which offer enhanced capabilities and functionalities.

Appendix A

ChatGPT Perception Survey

Please indicate your satisfaction with the ChatGPT learning environment in respect to your writing progress.

Statement	Disagree	Strongly disagree	Neither agree nor disagree	Agree	Strongly agree
ChatGPT has helped me improve my topic sentences					
ChatGPT has helped me improve my supporting sentences					
ChatGPT has helped me improve my concluding sentences					
ChatGPT has helped me improve my grammar					
ChatGPT has helped me to use verb tenses correctly					
ChatGPT has helped me to write sentences in the correct word order					
ChatGPT has helped me use personal pronouns correctly					
ChatGPT has helped me avoid making spelling errors					
ChatGPT has helped me correct my fragments and run-on sentences					
ChatGPT has helped me improve my use of capitalization					
ChatGPT has helped me improve my use of full stops					
Overall, ChatGPT has effectively improved my writing skills					
Using ChatGPT has been easy					
I have had a positive experience receiving corrective feedback from ChatGPT.					
I recommend using ChatGPT for written corrective feedback					
I Prefer receiving feedback from ChatGPT rather than from my instructor					
I Am aware of the ethical issues related to the use of ChatGPT, such as plagiarism					
I think the feedback I have received from ChatGPT is accurate					

Express your opinion about the experience of using ChatGPT to develop your writing skills. Do you find it useful or not? What are the advantages and disadvantages you have encountered in this experience?.

Acknowledgments

The authors would like to thank Dr Sanghoon Park for his valuable comments and all the participants who participated in this study.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Ethical Statement

Informed Consent

Informed consent was obtained from all participants, and all personal information was anonymized to maintain confidentiality. The participants were informed of their right to withdraw from the study at any time.

ORCID iD

Amal M. Barzanji https://orcid.org/0009-0007-3918-1573

References

Ahmed, M. (2023). ChatGPT and the EFL classroom: Supplement or substitute in Saudi Arabia's eastern region. *Information Sciences Letters*, *12*(7), 2727–2734. https://doi.org/10.18576/isl/120704

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://doi.org/10.24093/awej/call9.1

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing*, *17*(2), 102–118. https://doi.org/10.1016/j.jslw.2007.11.004

Bitchener, J., & Ferris, D. R. (2012). *Written corrective feedback in second language acquisition and writing*: Routledge.

Bitchener, J., & Knoch, U. (2008). The value of written corrective feedback for migrant and international students. *Language Teaching Research*, *12*(3), 409–431. https://doi.org/10.1177/1362168808089924

Bitchener, J., & Knoch, U. (2010a). The contribution of written corrective feedback to language development: A ten month investigation. *Applied Linguistics*, *31*(2), 193–214. https://doi.org/10.1093/applin/amp016

Bitchener, J., & Knoch, U. (2010b). Raising the linguistic accuracy level of advanced L2 writers with written corrective feedback. *Journal of Second Language Writing*, *19*(4), 207–217. https://doi.org/10.1016/j.jslw.2010.10.002

Bitchener, J., & Storch, N. (2016). *Written corrective feedback for L2 development*. Multilingual Matters.

Bonner, E., Lege, R., & Frazier, E. (2023). Large language model-based artificial intelligence in the language classroom: Practical ideas for teaching. *Teaching English with Technology*, *2023*(1), 23–41. https://doi.org/10.56297/Bkam1691/Wieo1749

Boudouaia, A., Mouas, S., & Kouider, B. (2024). A study on ChatGPT-4 as an innovative approach to enhancing English as a foreign language writing learning. *Journal of Educational Computing Research*, *62*(6), 1509–1537. https://doi.org/10.1177/07356331241247465

Chan, C., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *Int J Educ Technol High Educ*, *20*(1), 43. https://doi.org/10.1186/s41239-023-00411-8

Creswell, J. (2008). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (3rd ed.): Pearson/Merrill Prentice Hall.

Creswell, J. (2014). The selection of a research approach. In J. Creswell (Ed.), *Research design: Qualitative, quantitative, and mixed methods approaches* (pp. 3–23). Sage.

Eke, D. (2023). ChatGPT and the rise of generative AI: Threat to academic integrity? *Journal of Responsible Technology*, *13*, Article 100060. https://doi.org/10.1016/j.jrt.2023.100060

Ellis, N. C., Simpson Vlach, R. I. T. A., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *Tesol Quarterly*, *42*(3), 375–396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

Ferris, D. (1999). The case for grammar correction in L2 writing classes: A response to truscott (1996). *Journal of Second Language Writing*, *8*(1), 1–11. https://doi.org/10.1016/S1060-3743(99)80110-6

Ferris, D. (2008). Students must learn to correct all their writing errors. In J. Reid (Ed.), *Writing myths: Applying Second language research to classroom teaching* (pp. 90–114): University of Michigan Press.

Ferris, D. (2011). *Treatment of error in second language student writing* (2nd ed.): University of Michigan Press.

Fosnot, C. (1996). Constructivism: A psychological theory of learning. In C. T. Fosnot (Ed.), *In constructivism: Theory, perspectives and practice* (pp. 8–33): Teachers' College Press.

Fu, Q. K., Zou, D., Xie, H., & Cheng, G. (2022). A review of AWE feedback: Types, learning outcomes, and implications. *Computer Assisted Language Learning*, *37*(1–2), 179–221. https://doi.org/10.1080/09588221.2022.2033787

Guo, K., & Wang, D. (2023). 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

Han, J., Yoo, H., Kim, Y., Myung, J., Kim, M., Lim, H., Kim, J., Lee, T., Hong, H., Ahn, S., & Oh, A. (2023). Recipe: How to integrate ChatGPT into EFL writing education. In *Proceedings of*

the tenth ACM conference on learning @ scale (pp. 416–420). Association for Computing Machinery. https://doi.org/10.1145/3573051.3596200

IBM. (2024). What are AI hallucinations? https://www.ibm.com/topics/ai-hallucinations

Kim, S., Shim, J., & Shim, J. (2023). A study on the utilization of OpenAI ChatGPT as a second language learning tool. *Journal of Multimedia Information System*, *10*(1), 79–88. https://doi.org/10.33851/JMIS.2023.10.1.79

Lantolf, J., & Thorne, S. (2006). *Sociocultural theory and the genesis of second language development*: Oxford University Press.

Lim, W., Gunasekara, A., Pallant, J., Pallant, J., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *International Journal of Management in Education*, *21*(2), Article 100790. https://doi.org/10.1016/j.ijme.2023.100790

Lock, S. (2022, December 6). What is AI chatbot phenomenon ChatGPT and could it replace humans. *The Guardian*. https://rb.gy/nisk56

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments*, *11*(1), 9. https://doi.org/10.1186/s40561-024-00295-9

Masoudi, H. (2024). Effectiveness of ChatGPT in improving English writing proficiency among Non-native English speakers. *International Journal of Educational Sciences and Arts*, *3*(4), 62–84. https://doi.org/10.59992/IJESA.2024.v3n4p2

Mathur, M., & Mahapatra, S. (2022). Impact of ePortfolio assessment as an instructional strategy on students' academic speaking skills: An experimental study. *CALL-EJ*, *23*(3), 1–23.

Mohsen, M. A. (2022). Computer-mediated corrective feedback to improve L2 writing skills: A meta-analysis. *Journal of Educational Computing Research*, *60*(5), 1253–1276. https://doi.org/10.1177/07356331211064066

OpenAI. (2024). Introduction. https://platform.openai.com/docs/models

Peres, R., Schreier, M., Schweidel, D., & Sorescu, A. (2023). On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice. *International Journal of Research in Marketing*, *40*(2), 269–275. https://doi.org/10.1016/j.ijresmar.2023.03.001

Sheen, Y. (2007). The effect of focused written corrective feedback and language aptitude on ESL learners' acquisition of articles. *Tesol Quarterly*, *41*(2), 255–283. https://doi.org/10.1002/j.1545-7249.2007.tb00059.x

Sheen, Y., Wright, D., & Moldawa, A. (2009). Differential effects of focused and unfocused written correction on the accurate use of grammatical forms by adult ESL learners. *System*, *37*(4), 556–569. https://doi.org/10.1016/j.system.2009.09.002

Shintani, N., & Ellis, R. (2013). The comparative effect of direct written corrective feedback and metalinguistic explanation on learners' explicit and implicit knowledge of the English indefinite article. *Journal of Second Language Writing*, *22*(3), 286–306. https://doi.org/10.1016/j.jslw.2013.03.011

Shintani, N., Ellis, R., & Suzuki, W. (2014). Effects of written feedback and revision on learners' accuracy in using two English grammatical structures. *Language and learning*, *64*(1), 103–131. https://doi.org/10.1111/lang.12029

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*, Article 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing*, *19*, 51–65. https://doi.org/10.1016/j.asw.2013.11.007

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language and learning*, *46*(2), 327–369. https://doi.org/10.1111/j.1467-1770.1996.tb01238.x

von Glasersfeld, E. (1996). Introduction: Aspects of constructivism. In C. T. Fosnot (Ed.), *In Constructivism: Theory, perspectives and practice* (pp. 3–7): Teachers' College Press.

Wilson, J., & Andrada, G. N. (2016). Using automated feedback to improve writing quality: Opportunities and challenges. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 678–703): IGI Global.

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, *100*, 94–109. https://doi.org/10.1016/j.compedu.2016.05.004

Xiao, Y., & Zhi, Y. (2023). An exploratory study of EFL learners' use of ChatGPT for language learning tasks: Experience and perceptions. *Language*, *8*(3), 212. https://doi.org/10.3390/languages8030212

Yusuf, A., Pervin, N., & Román-González, M. (2024). Generative AI and the future of higher education: A threat to academic integrity or reformation? Evidence from multicultural perspectives. *Int J Educ Technol High Educ*, *21*(1), 21. https://doi.org/10.1186/s41239-024-00453-6

Zhai, N., & Ma, X. (2023). The effectiveness of automated writing evaluation on writing quality: A meta-analysis. *Journal of Educational Computing Research*, *61*(4), 875–900. https://doi.org/10.1177/07356331221127300

Author Biographies

Abrar H. Alsofyani is an assistant professor at the English Language Institute (ELI), King Abdulaziz University. She earned her B.A. in European Languages: English from King Abdulaziz University. In 2011, she was awarded an MA in ESL Curriculum and Instruction from Tennessee State University, USA, and in 2019, she completed her Ph.D. in Technology in Education and Second Language Acquisition from the University of South Florida, USA. Her research interests include ESL/EFL written corrective feedback, reading strategies and comprehension, collaborative learning, language pedagogy, artificial intelligence in education, and instructional technology.

Amal M. Barzanji is an assistant professor at Taiba University. She earned her Ph.D. in Technology in Education and Second Language Acquisition from the University of

South Florida (2021). Her Master's was in Teaching English to Speakers of Other Languages from the University of Central Florida (2016). She received her B.A. in English language from Taibah University (2007). Her research interests include L2 writing, error analysis, language transfer, interlanguage, written corrective feedback, multimodal feedback, multimodal composition, multi-word expressions, and technology-based instruction.