Examining the Efficacy of ChatGPT and Human-Derived Corrective Feedback in Addressing Grammatical Errors in Saudi EFL Students' Compositions

Fawaz Al Mahmud[1]

[1] Faculty of English Languages and Translation, Department of English Language, University of Jeddah, Jeddah, Saudi Arabia

Correspondence: Fawaz Al Mahmud, Faculty of English Languages and Translation, University of Jeddah, Jeddah, Saudi Arabia. E-mail: falaswad@uj.edu.sa

Received: July 18, 2025 Accepted: September 16, 2025 Online Published: October 21, 2025

doi:10.5430/wjel.v16n2p258 URL: https://doi.org/10.5430/wjel.v16n2p258

Abstract

Employing an explanatory mixed-methods approach, this study compared the affordances of ChatGPT's written corrective feedback (WCF) and human corrective feedback in fostering linguistic accuracy among Saudi EFL learners. Participants were 53 undergraduates at the University of Jeddah, divided into an experimental group (n = 26) and a control group (n = 27). Over eight weeks, both groups engaged in two 1.5-hour writing sessions per week. The experimental group received WCF from ChatGPT, while the control group received human WCF. Pretests and posttests assessed participants' grammatical competence before and after the intervention. Error-flagging accuracy was evaluated by a panel of experienced university teachers of English. Semi-structured interviews with six experimental-group participants explored perceptions of ChatGPT. Quantitative findings indicated that ChatGPT was a more effective source of WCF than human feedback, as the experimental group significantly outperformed the control group in the posttest. However, human feedback was slightly more accurate in error-flagging. Qualitative results revealed that most students valued ChatGPT for its precise error identification, though some considered human feedback clearer in explanation. The study concludes that ChatGPT can serve as a valuable supplementary tool for corrective feedback in Saudi EFL writing.

Keywords: automated written corrective feedback, Saudi EFL learner, academic writing, linguistic accuracy, surface errors

1. Introduction

Writing and corrective feedback complement each other in facilitating EFL proficiency development. By composing different kinds of documents in English, students are initiated into noticing and reflecting upon their own metalinguistic competence, which gradually refines their written-English repertoire. However, valuable as writing is for L2 writing development, this output must receive the instructor's written corrective feedback (henceforth WCF) in order for learners to notice pitfalls in their compositions. WCF is crucial to EFL writing because it helps learners minimize errors and write better English. (Anderson & Ayaawan, 2023; Butterfuss et al., 2022; Panadero et al., 2023; Bitchener and Ferris, 2012). An effective WCF does so by telling students what they need to review and revise and how to carry out that revision. But human WCF becomes a taxing enterprise when the instructor is dealing with large classes (Graham, 2019) where he has to read a multitude of drafts painstakingly and provide timely WCF on them. Consequently, instructors are either not able to attend to every single document or necessarily have to limit and dilute their WCF. Therefore, taking steps to lighten the load on instructors to be the sole providers of WCF will positively impact writing instruction. To this end, automated writing tools, such as Grammarly and ChatGPT, have emerged as capable academic auxiliaries for learners as well as instructors.

AI-driven writing tools are software applications that can compose and evaluate texts (Lund & Wang, 2023). Given their increasing use in writing and writing assessment, automated writing tools have been the focus of many studies (Chapelle, Cotos, & Lee, 2015; Radford, Wu, Child, Luan, Amodei & Sutskever, 2019). These tools have been found to raise learners' motivation and help them become better writers (Graham, Hebert, & Harris, 2015; Wilson & MacArthur, 2024). However, it has also been found that automated WCF sometimes fails to match the accuracy of human WCF and at other times is confusing to the learner (Wilson & MacArthur, 2024). Moreover, fine-tuning such tools for use in academic settings has been found to take a lot of time for the reason that most of the automated writing evaluation tools require extensive training on hundreds of well-written, human-rated essays on the same topic before they, that is, the digital writing tools, can be used reliably to evaluate student texts.

However, the release of ChatGPT, in November 2022, has changed the situation for the better. Since it does not require prior training on authentic human texts in order to compose or provide feedback, ChatGPT has become a much sought-after tool for students and instructors alike. Compared with its predecessors, ChatGPT has been reported to be more efficient in spotting, elucidating, and fixing surface errors in texts.

2. Literature Review

2.1 Automated WCF and EFL Writing

Numerous studies have examined how automated WCF impacts the linguistic accuracy of EFL students and have found that it helps learners reduce surface errors (Dizon & Gayed, 2021; Schmidt-Fajlik, 2023). This kind of research has tended to focus on two areas of investigation. The first area of inquiry has reported automated WCF-induced gains in grammatical accuracy across multiple writing samples produced by the same EFL learners (Thi & Nikolov, 2022; Liao, 2016). Liao (2016), studying automated WCF as learners' grammatical-accuracy enhancer, first isolated error types in writings of 66 Taiwanese-university EFL students and subsequently measured grammatical errors in their writing intermittently over a nine-week teaching regime, using automated WCF. This feedback was subjected to descriptive analysis and paired-samples t tests. Results indicated that automated WCF helped students reduce grammatical errors. In another study, Ghaemi and Bayati (2021), seeking to see whether WhiteSmoke Writing Software helped EFL learners reduce surface errors in their writing, found that the Experimental group, which used WhiteSmoke, made fewer errors in their posttest compositions compared with the Control group. The conclusion was that WhiteSmoke benefited learners. Why automated WCF worked for the participants in the above-mentioned studies is not difficult to explain. While one reason may be that these automated tools provide feedback immediately, the other reason might be that automated WCF is usually more comprehensive than human WCF.

While the studies reviewed above all examined the role of automated WCF in mitigating surface errors in texts by the same EFL students, another line of inquiry compared automated WCF with human WCF; some studies have concluded that automated WCF is more effective than human WCF in helping students minimize surface errors (Wang & Li, 2021; Wang, 2011). For instance, Wang, S. (2011) inquired into the role of Writing Roadmap (WRM), an automated writing-evaluation tool, in helping L2 students write with greater grammatical accuracy. The findings demonstrated that learners using WRM feedback outperformed those working with human WCF in linguistic accuracy. This, interestingly, was the case for both the WRM-evaluated essays and the teacher-evaluated essays. In a related study, Han and Sari (2024) divided the EFL students of a Turkish university into two groups. While one group received both automated and human WCF on its written output, the other group received human WCF only. The analysis of post-feedback writing samples revealed that although learners in both groups committed fewer errors in grammar compared with the errors, they made in pre-feedback writing, the group that received the combined feedback from software and humans wrote with greater grammatical accuracy than the group that received human WCF only.

2.2 ChatGPT and EFL Writing

The efficacy of ChatGPT as a source of WCF is still an emerging domain in EFL writing research. Nonetheless, the facilitative role of ChatGPT in L2 writing has been acknowledged in the literature (Petersen & Jabbari, 2024; Barrot, 2024). While some researchers have praised ChatGPT for its interactivity, task-oriented content generation, and resourcefulness (Jeon & Lee, 2023), others have appreciated this AWE tool for its ability to provide instant, comprehensive, and predominantly-accurate feedback on writing, especially on linguistic accuracy (Mizumoto et al., 2024). Inquiries into ChatGPT's affordances as a WCF tool have focused on two main areas: the first area comprises studies comparing the WCF performance of ChatGPT with that of human evaluators (Song & Song, 2023; Fokides & Peristeraki, 2024); the second area comprises research evaluating the WCF performance of ChatGPT only (Mahapatra, 2024).

A pioneering study of ChatGPT as a WCF tool was Pfau et al. (2023). They had ChatGPT catch and classify grammar errors in 100 essays by EFL learners. Subsequently, human experts examined those essays and ChatGPT's WCF to ascertain how precisely ChatGPT's performance was. Findings demonstrated a significant correlation between ChatGPT’s error flagging and that of human experts. Since it showed ChatGPT to be highly accurate in error flagging, this study inspired other studies to explore the role of ChatGPT as a WCF tool. However, Pfau's study had some limitations. For example, it chose ChatGPT as the sole representative of AWCF, omitting Grammarly, a well-known AWE tool. Had Pfau evaluated ChatGPT and Grammarly side by side, a richer picture of ChatGPT's WCF performance would have emerged.

While Pfau et al. were focused on comparing the error-detection and error-classification capabilities of the chatbot and the human evaluators, other studies explored how ChatGPT's WCF helped L2 learners write grammatically-accurate prose. Song and Song (2023) in their mixed methods study evaluated the role of ChatGPT in enhancing writing skills of Chinese EFL learners. Results showed that the experimental group, working with ChatGPT instruction, outperformed the control group, which received traditional instruction. Another study, by Fokides & Peristeraki (2024), reached similar conclusions. Examining the affordances of ChatGPT's WCF on school learners' English and Greek essays, Fokides & Peristeraki (2024) found that for English essays, ChatGPT's WCF outperformed human experts by spotting more errors and providing richer feedback. These findings augmented the conclusion of previous studies that ChatGPT was a potent tool for WCF, on par with human WCF.

Since the findings from the studies reviewed above show that usually ChatGPT's error-flagging is accurate and reformulations are effective, a favorable picture of ChatGPT as a viable supplement or alternative to human WCF emerges. However, there still are gaps in literature. For instance, to the best of the author's knowledge, no extant study has addressed the question of ChatGPT's role in facilitating linguistic accuracy in conjunction with the question of the error-flagging accuracy of ChatGPT itself. Additionally, at the time of the writing of this study, there are few research papers devoted to an exploration of the potential of ChatGPT as a source of WCF for Saudi EFL learners. The present study attempts to fill these gaps. The study will be guided by the following research questions.

RQ 1) How do ChatGPT and human counsels compare on the effectiveness of feedback function?

RQ 2) How do ChatGPT and human counsels compare on the accuracy of the feedback?

RQ 3) How do the Saudi EFL learners feel about the efficacy of ChatGPT as a tool for WCF?

This study has a two-pronged significance. First, while there have been many inquiries into the role of WCF in EFL writing development, most of these works have focused either on the human WCF or on the automated WCF. Few researchers have opted to compare the relative effectiveness of the two different sources of WCF. A comparative study, therefore, is timely and valuable as it will help the stakeholders decide whether or not automated WCF should be employed as an alternative to human WCF at least in certain specific contexts.

The remainder of this paper is organized as follows. The 'Method' section describes the study participants, research instruments, and methods of data analysis. The results are then furnished and discussed, followed by recommendations for stakeholders and the limitations of the study. The paper concludes by recapitulating the major findings of this study.

3. Method

3.1 Research Design

Using the explanatory sequential mixed-method (Creswell, 2018), this study compared ChatGPT WCF and human WCF in helping Saudi EFL learners write grammatically accurate prose. The learners' perception of the efficacy of ChatGPT's and human WCF was also explored. For RQ 1, the independent variables were the feedback conditions: ChatGPT's WCF (experimental group); human WCF (control group). The dependent variable was participants' posttest written-English performance in terms of grammatical accuracy. For RQ 2, both the ChatGPT feedback and the instructor feedback were compared for error-flagging accuracy by experienced teachers of English. For RQ 3, six participants were interviewed to foreground their perceptions about how good, or otherwise, the ChatGPT was as a facilitator of writing accuracy.

3.2 Participants

The participants were 53 Saudi undergraduate EFL learners at the University of Jeddah divided into the experimental group (n=26) and the control group (n=27). The participants took an academic writing course at the University of Jeddah, Saudi Arabia.

3.3 Research Instruments

The pretest and the posttest were used to collect the participants' writing samples. ChatGPT and human evaluators were the sources of feedback. As regards the details and credentials of human evaluators, four Saudi university teachers of English agreed to the author's request to examine the participating students' written work, evaluate grammatical accuracy, and provide corrective feedback. Each of these evaluators had a PhD in English with 5-8 years of experience teaching EFL to Saudi college students, besides 4-8 relevant research publications in remedial grammar and academic writing. To measure the grammatical accuracy of students' writing, the evaluators all agreed to use a formula employed by Soltanpour and Valizadeh (2018): [total number of grammatical errors/total number of words] × 100 to calculate the mean number of errors per 100 words of the essay.

3.4 Research Procedure

After assignment to the experimental or control group, all 53 participants took a pretest designed to assess their grammatical proficiency at the start of the study. In this test, each participant wrote a short essay, which was assessed and marked for grammatical accuracy by a human evaluator. Each participant received a score out of 100, calculated on the basis of the mean number of errors per hundred words.

Over the next eight weeks, both groups engaged in two writing sessions per week, with each session lasting 1.5 hours. In every session, participants wrote short essays and received corrective feedback. The experimental group received written corrective feedback (WCF) from ChatGPT once per session, while the control group received human WCF once per session. This cycle of writing, receiving feedback, and revising continued across all sessions.

In the ninth week, all participants completed a posttest by writing another short essay. These essays were assessed for grammatical accuracy by independent human raters using the same scoring procedure as the pretest. In judging linguistic accuracy, Huddleston & Pullum (2002) and Quirk et al. (1985) served as reference yardsticks. After the posttest, six participants from the experimental group were purposively selected for semi-structured interviews exploring their perceptions of ChatGPT as a source of WCF.

To examine the reliability of ChatGPT and human WCF, an additional interrater check was conducted on a representative subset of 12 essays (≈23% of the dataset). This number was chosen to provide a manageable yet balanced sample across proficiency levels in both groups, allowing a consistency check without re-evaluating the entire corpus. The two raters independently provided corrective feedback, and their agreement was used to assess the consistency of ChatGPT feedback with human feedback.

3.5 Data Analysis

The data were analyzed quantitatively as well as qualitatively. The quantitative data was analyzed in the SPSS. To obtain the general characteristics of the distribution of scores, means and standard deviations for each item were calculated. As for the qualitative data, error analysis and thematic analysis were employed. Error analysis was used to determine the relative accuracy of ChatGPT's feedback and human feedback. Finally, the students' perceptions of ChatGPT as a feedback tool were ascertained through thematic analysis. First the interview transcripts were read to obtain a general sense of the interview. Next, the transcripts were coded manually, breaking the text

down into manageable units and labelling them to develop descriptions and themes (Saldana, 2021).

4. Results

4.1 Quantitative Results

To examine the relative impact of ChatGPT and human WCF in fostering writing development of Saudi EFL students, 53 Saudi undergraduate EFL learners at the University of Jeddah were selected and divided into the experimental group, (n=26), which received ChatGPT's WCF, and the control group (n=27), which received human WCF.

Table 1 below shows the mean, standard deviation, and t- test for two independent samples, to see if there is any statistically significant difference between the mean scores of the control group and the mean scores of the experimental group in their levels of writing skills at the start of the study.

Table 1. The mean, standard deviation, and t-test in writing skills at the premeasurement

Test	Group	n	Mean	Std. Deviation	t-test	P-value
Pre-test writing skills	Experimental group	26	54.61	13.10	0.549	0.586
	Control group	27	52.77	11.10		

Table 1 above identifies that the value of the T-test to test writing skills is statistically insignificant because the P- value associated with the test is higher than the level of statistical significance ($\alpha = 0.05$). So, one should accept the hypothesis that there is no statistically significant difference between the average scores of the participants of control and experimental groups on the pre- test. It, then, follows that the scores of the control and experimental groups are equal before the start of this research, which means the variable WCF from ChatGPT will be the only difference between the two groups after removing all confounding variables.

Additionally, the mean, standard deviation, and two independent t- test were performed to assess whether there is a significant difference between the mean scores of the control group and experimental group or not, in writing skills in the post measurement.

Table 2. the mean, standard deviation, and t- test for writing skills in the post measurement

Test	group	n	mean	Standard deviation	t- test	P- value	Eta square
Post-test writing skills	Experimental group	26	66.15	11.34	3.14	0.003	0.68
	Control group	27	56.48	11.07			

Table 2 above shows that there is a statistically significant difference at the level of significance ($\alpha = 0.01$) between the mean scores of the experimental and control group in writing skills.

Writing skills tests and difference was in support of the experimental group (writing corrective feedback from ChatGPT) comparing to the control group (human-derived writing corrective feedback), where the mean of their writing skills test (66.15) was higher than that of the control group (56.48) and statistically significant ($t = 3.14$, $P < 0.001$), which implies the statistical significance of the effectiveness of applying the method of getting writing corrective feedback from ChatGPT in remediating writing skills. The effect size (eta square $\eta 2$) for getting writing corrective feedback from ChatGPT as learning method in writing skills was (0.68), this means that 68% of the variance in writing skills is associated with applying the method of getting writing corrective feedback from ChatGPT in remediating writing skills.

4.2 Qualitative Results

Error Detection Accuracy: ChatGPT Feedback versus Human Feedback

This subsection furnishes the qualitative findings on the error-detection accuracy of ChatGPT's and human WCF. A total of 889 errors in grammar, punctuation, and word choice were correctly identified by the human rater (486 errors) plus ChatGPT (403 errors) across a sample of fifty-three 270-300 word essays composed by a total of 53 students (27 control-group essays plus 26 experimental-group essays). While three surface errors (all related to grammar) went undetected (false negatives) in human feedback, seven surface errors (one related to lexis and six related to grammar) went undetected (false negatives) in ChatGPT's feedback. Thus, a total of 899 surface errors were actually there in 53 essays (486+03=489 errors in 27 control-group essays and 403+07=410 errors in 26 experimental-group essays). For both the experimental and the control groups, the most frequently-occurring errors were errors in grammar, followed by errors in punctuation and then by errors in word choice.

As regards the error-detection accuracy, the human rater outperformed the ChatGPT by a narrow margin. Here, it should also be stated that overall ChatGPT did a brilliant job of detecting surface errors in student writing samples. The table below summarizes the relative performance of the human rater and ChatGPT as tools for flagging surface errors in the writing samples produced by the participants of this study.

Table 3. The relative accuracy of error detection by the human rater and the ChatGPT

Error type	*Total errors*	*Correctly identified by human rater*	*Correctly identified by ChatGPT*	*Missed by human rater*	*Missed by ChatGPT*
Lexical Errors	96	59	36	00	01
Grammatical Errors	640	338	293	03	06
Punctuation Errors	163	89	74	00	00

Total errors (899): the sum of errors correctly identified in the written corrective feedback by human instructor (486) and ChatGPT (403) plus the sum of errors missed by human instructor (03) and ChatGPT (07)

As can be inferred from the tabular data above, both the human rater and ChatGPT correctly identified errors in most instances, with the percentage of false negatives under five percent for both.

Errors Correctly Identified By Human Rater: The following types of errors were correctly identified by the human rater on every single occasion: the use of a wrong word; the use of incorrect forms of nouns and verbs; the unnecessary use of the definite article; the use of incorrect adverb; squinting modifiers; the use of wrong preposition; faulty subject-verb agreement involving linking verbs, compound subjects, separated subjects, and collective-noun subjects; verb-tense errors; comma splice; fused sentence; and sentence fragments. A few excerpts from the control-group writing samples will illustrate the point more fully. The sentences below have all been excerpted from the control-group student essays composed in response to the following prompt: "Should Phoning While Driving Be Banned?"

The use of a wrong word

Phones have become a principle tool for communication in our lives.

(CGE9: Control Group Essay 9).

The human rater correctly identified the error of the use of wrong word 'principle' and suggested the right word 'principal'.

The use of incorrect noun form

For example, phone help us connect with family and friends. (CGE3)

The human rater correctly identified the error of the use of incorrect noun form 'phone' and suggested the correct noun form 'phones'.

The unnecessary use of the definite article

Nevertheless, you have to get rid of the things that may grab your attention, including the phones as they are the center of distraction. (CGE20)

The human rater correctly identified the unnecessary use of the definite article and recommended that it be deleted.

The use of incorrect adverb

When on the road, which of the two is the most pressing concern, focus on driving or keeping in touch? (CGE21)

The human rater correctly identified the incorrect use of the adverb 'most' and recommended that it be replaced with the correct adverb 'more'.

Squinting modifier

Those who indulge in phoning while driving often will meet an accident. (CGE11)

The human rater correctly identified the squinting modifier 'often' and suggested the following revision: Those who often indulge in phoning while driving will meet an accident.

The use of wrong preposition

In the end remember that if you are safe you will make money to buy food to your family. (CGE13)

The human rater correctly flagged the wrong preposition 'to' and recommended that it be replaced with 'for'.

Faulty Subject-Verb Agreement

Ok, so phoning while driving plus the chances of hitting another car is the focus of my essay. (CGE10)

The human rater correctly flagged the error 'is' and suggested the revision: 'are'.

Faulty Verb Tense

I am hearing about this idea of banning the use of phones when you are driving for months now. (CGE12)

The human rater correctly flagged the error 'am' and suggested the revision 'have been'.

Comma Splice

So we see a lot of such drivers, they always use phone. (CGE20)

The human rater correctly flagged the error ',' and suggested the revision: ";".

Sentence Fragments

The main argument for banning phone use while driving is safety. Because drivers using phones, even hands free, are much more likely to get in a crash. (CGE1)

The human rater correctly flagged the error 'fragment' and suggested the revision:

The main argument for banning phone use while driving is safety because drivers using phones, even hands free, are much more likely to get in a crash.

Errors Missed By Human Rater: As has been stated earlier, the human rater flagged errors accurately most of the time. However, he too failed to detect four errors in grammar and two errors in punctuation in the CGS writing samples. The following erroneous sentences went undetected.

Faulty use of Hypothetical Conditional

In conclusion, I will say this: If I have the authority, I would ban phoning while driving today.

Revised: If I *had* the authority, I would ban phoning while driving today.

Faulty inversion in an indirect question

After waiting a while for the car in front of my car to give me way, I felt like yelling this: Brother, can I please ask you just why are you not listening?

Revised:

After waiting a while for the car in front of my car to give me way, I felt like yelling this: Brother, can I please ask you just why you *are* not listening?

Faulty use of the Verbal

Although he was not hurt in the accident, he denied to take responsibility for the accident.

Revised: Although he was not hurt in the accident, he denied taking responsibility for the accident.

Errors Correctly Identified By ChatGPT.

Our findings reveal that ChatGPT was almost as accurate as the human rater in flagging errors. It accurately identified surface errors across several domains, including article usage (missing article; unnecessary article, wrong article); subject-Verb agreement (straightforward concord; separated subjects; collective nouns as subjects; indefinite pronouns as subjects); preposition usage (missing preposition; unnecessary preposition; wrong preposition) verb usage (wrong tense; wrong auxiliary verb); adjectives (order of adjectives; gradable/non-gradable adjectives; attributive/predicative adjectives); modifier placement; comma usage; and sentence fragments.

One instance of ChatGPT's corrective feedback on an experimental-group essay (EGE4) will illustrate the point more fully. The essay was written in response to the following prompt: "Should Phoning While Driving Be Banned?" The original essay, reproduced verbatim, appears below, followed by the ChatGPT's corrective feedback on it.

The Essay (EGE4)

There is a big debate about whether people should be allowed to use phone while driving. Both sides have good points.
The main reason why phone use should be banned while driving is safety. A lot of research shows that if drivers use phone and if they drive at same time they may crash their car. Phone distract drivers from the important, visual, physical, and mental tasks of driving safely. This can make drivers slower in reacting and less aware of there surroundings. This greatly increases the chance of accident.
Given these safety risks, many people think that government not aware but they did. Because of this we see many accidents. Because of this many drivers do not have any solution for using phone while driving but actually we have many of them. I mean many of the solutions. One thing is using blotooth or wearing a head phone. You can say this is the kind of solution. They should avoid using phone to not cause an accident.

However, we see a lot of drivers do not give any care of that, so that is way we put rules. We hear everyday people get an accident and some them died because of that. The government put violation who use phone while driving and it cost about 1000 Saudi riyals, or more than that. So you must keep follow the rules and instruction.
Overall, this what we should and what we need. The important things is be careful when you driving, and stop using phone while driving. You have a people life with your hand. Be smart you do not want risk your life because using phone. Just wear seatbelt, be focus in the road please for your own life. That is what we should do to live your life safely.

ChatGPT was asked to provide WCF on the above essay. The feedback appears below.

ChatGPT Feedback

Sure! Let's break down the text for errors in grammar, lexis, and punctuation, and then revise each erroneous element.

Errors Identified

1. Grammar Errors:
    - "phone" should be pluralized as "phones."
    - "if drivers use phone and if they drive at same time" should include articles ("the") and have parallel structure.
    - "there surroundings" should be "their surroundings."
    - "the chance of accident" should be "the chance of an accident."
    - "government not aware but they did" is unclear and grammatically incorrect.
    - "the kind of solution" is awkwardly phrased.
    - "that is way we put rules" should be "that is why we put rules."
    - "some them died" should be "some of them died."
    - "put violation who use phone" is unclear and should be rephrased.
    - "keep follow the rules" should be "keep following the rules."
    - "the important things is be careful when you driving" should be "the important thing is to be careful when you are driving."
    - "you have a people life with your hand" is awkward and unclear.
    - "you do not want risk your life" should be "you do not want to risk your life."
    - "be focus in the road" should be "be focused on the road."
    - "to live your life safely" is awkward; it would be better to phrase it differently.
2. Lexical Errors:
    - "blotooth" should be "Bluetooth."
    - "head phone" should be "headphones."
    - "be focus in the road" should use "focused."
    - "you have a people life" should be rephrased for clarity.
3. Punctuation Errors:
    - Missing commas for clarity in several sentences.
    - "and stop using phone while driving" lacks a comma before "and" in a compound sentence.

As can be deduced from the ChatGPT feedback on EGE4 and as was found in the ChatGPT feedback on other experimental-group essays, the AI tool correctly identified surface errors in the following domains: Article usage (missing article; unnecessary article, wrong article); Subject-Verb Agreement (straight-forward concord; separated subjects; collective nouns as subjects; indefinite pronouns as subjects); preposition usage (missing preposition; unnecessary preposition; wrong preposition) verb usage (wrong tense; wrong auxiliary verb); adjectives (order of adjectives; gradable/non-gradable adjectives; attributive/predicative adjectives); modifier placement; sentence fragments.

Errors Missed By ChatGPT

It must be stated upfront that ChatGPT flagged errors accurately in most instances and that the few errors that ChatGPT could not detect were minor lapses and did not impede communication in any noteworthy way. Nonetheless, those errors, along with the corresponding corrections, are listed below.

The use of wrong word

At the end for me there is not much distinction between a person who uses phone while driving and a criminal.

Revised: At the end for me there is not much *difference* between a person who uses phone while driving and a criminal.

The use of past-time adverb with 'will/shall + base form of the verb'

And when I asked Nawaf, the guy who hit my car because he was on the phone, to pay me money, he simply said this: "No problem, brother. I will get back to you after one week."

Revised: And when I asked Nawaf, the guy who hit my car because he was on the phone, to pay me money, he simply said this: "No problem, brother. I will get back to you in one week."

The use of 'According to me' construction

Well, according to me, phoning while driving should be banned because you have no right to create problems for others.

Revised: Well, in my opinion, phoning while driving should be banned because you have no right to create problems for others.

Omission of contextual 'the'

The car stops and the guy who was warned by the police not to use phone while driving walks into the restaurant speaking over phone. Now he is having his meal. He casually turns to the person sitting next to him and asks, 'Could you please pass me salt?"

Revised: The car stops and the guy who was warned by the police not to use phone while driving walks into the restaurant speaking over phone. Now he is having his meal. He casually turns to the person sitting next to him and asks, 'Could you please pass me the salt?"

Stative verb used in the progressive

And now I am clearly hearing the loud argument between those involved in the accident.

Revised: And now I can clearly hear the loud argument between those involved in the accident.

Unnecessary use of preposition

I love doing three things: driving fast cars, using latest smartphones and climbing on mountains.

Revised: I love doing three things: driving fast cars, using latest smartphones and climbing mountains.

Vague Pronoun Reference

I was so busy talking over the phone, and I did not see the roadside electric pole, and my car crashed into the roadside electric pole. But Alhumdolillah it was not much damaged.

Revised: I was so busy talking over the phone, and I did not see the roadside electric pole, and my car crashed into the roadside electric pole. But Alhumdolillah *the car* was not much damaged.

Students' Perceptions About ChatGPT As A Facilitator of Writing Accuracy

To address RQ 3, interview transcripts of six participants were analysed. The analysis indicated that the participants felt positively about the corrective and facilitative role of ChatGPT in their writing endeavors. One interviewee, for example, had this to say about the efficacy of ChatGPT as a digital facilitator of grammatically-correct writing.

"I think I am having a very good experience with ChatGPT. The reason is that as soon as I give it my English work and ask it to tell me what is wrong in grammar or choice of words or punctuation, it gives me a detailed list of my errors. Not only this, ChatGPT also tells me why what is wrong is wrong. And then it changes my writing so that there are no errors in it. So it is fantastic!"

While this interviewee was exuberant about ChatGPT's error-spotting and fixing capabilities, some others were more focused and substantial in expressing views about the strength of ChatGPT a source of corrective feedback on their writing.

Describing ChatGPT's precise error spotting and clear descriptions of what exactly was wrong with an erroneous sentence, one interviewee made the following remarks:

"Since the time I started using ChatGPT as a source of corrective feedback, I must say I have learned a lot. And ChatGPT is really impressive. Previously, I would commit grammatical blunders, I would write such almost garbled sentences as the

> following: 'I am go'; 'In 2020 3308 were killed for distracted driving and using phone'; 'We can solve these problems without banning phone by working on it.'; Every year a lot die because of phone , and car accidents are the main reason.' But ChatGPT caught these errors and many other errors and clearly explained what the problem was in each case before going on to revise the sentences."

If the above-mentioned interviewees were all praise for ChatGPT's corrective feedback capabilities, the following interviewee was a bit equivocal in his perception of ChatGPT. In talking about ChatGPT as a provider of corrective feedback, he compared ChatGPT to human evaluator and made the following observation:

> "Well, in my experience ChatGPT is good at spotting errors, but sometimes its explanation of that error is unclear. For example, once ChatGPT told me that I should not write 'We all live in the society." Instead I should write 'We all live in society.'" But the explanation was not very clear. ChatGPT said it was a natural way to drop 'the' in that context. But when I asked my teacher, he clarified, saying that since I was not talking about any particular society known to both me and my reader, 'the' was not needed. So I would say ChatGPT is a good error spotter, but our teachers are better explainers."

As can be seen, the perception above is not in tune with the quantitative findings, which show that, compared to humans, ChatGPT is a more effective tool for helping students minimize surface errors in their compositions. This apparent discrepancy can be accounted for by considering a few inherent limitations of individual views as a form of qualitative data. To begin with, how a participant responds to and evaluates a tool depends considerably upon his or her own knowledge of and skills with that tool. For example, someone with an in-depth knowledge of English grammar might find a particular explanation, by ChatGPT or a human being, of a particular language point inadequate or unsatisfactory because he or she was expecting a detailed substantiation of why that specific linguistic construction should count as erroneous. Such a student would not be satisfied with error-labeling and correction alone. In contrast, a student with poor proficiency in English grammar will go on to praise any feedback that does just two things: flags error and suggests revision. Besides individual proficiency, individual expectation also shapes a participant's perspective. For instance, two learners may have different sets of linguistic challenges, abilities, and expectations. For one learner, using verbs and adverbs correctly may be the biggest challenge; for another one, the biggest challenge may be articles and prepositions. Thus, while one learner would expect ChatGPT to provide the most thorough explanations of his favorite set of errors, the other would expect his favorite errors to receive maximum attention. Thus, different learners will evaluate a learning tool, such as ChatGPT, with reference to their own expectations. It can, then, be said that considering the subjective nature of qualitative findings helps us reconcile the difference between quantitative and qualitative results.

Thus, it can be inferred from the foregoing examples of the interviewees' perceptions about ChatGPT that they unanimously acknowledged the efficacy of the Chat bot in helping Saudi students reduce surface errors in their writing. While the interviewees were not unanimous about the explanatory capabilities of ChatGPT, they all agreed that ChatGPT was an effective tool for identifying surface errors in Saudi student compositions.

5. Discussion

The aim of this investigation was to ascertain the relative efficacy of ChatGPT and human corrective feedback on surface errors in the writings of Saudi students and to fathom those students' perceptions of ChatGPT as a source of feedback on surface errors and facilitator of correct writing. The quantitative data were collected through a pretest and a posttest. The qualitative data were collected through the analysis of human and ChatGPT's feedback and through the semi-structured interviews of six participants. The data were analyzed statistically, grammatically and thematically. In this section, the discussion returns to the research questions and attempts to synthesize the findings of the present study, contextualizing them in relation to relevant extant literature.

RQ 1: How do ChatGPT and human feedback compare on the effectiveness of feedback function?

The first research question addressed was essentially the following: Which one is a greater help to Saudi students, ChatGPT's corrective feedback or human corrective feedback? The findings revealed that ChatGPT's feedback was noticeably more effective than human feedback. The quantitative results demonstrated that both the experimental group receiving feedback from ChatGPT and the control group receiving feedback from human instructors started with similar levels of writing skills. This was established through the pretest results, where no statistically significant difference was found between the groups ($t = 0.549$, $p = 0.586$). However, in the posttest, the experimental group outscored the control group by a significant margin, with a mean score of 66.15 compared to 56.48 for the control group ($t = 3.14$, $p = 0.003$). The significant enhancement in the linguistic proficiency of the experimental group suggests that ChatGPT was a more effective source of feedback compared with human corrective feedback. These findings are in tune with a number of previous studies which conclude that automated written corrective feedback on surface errors is as effective as or is more effective than human-derived written corrective feedback (Wang & Li, 2021; Mahapatra, 2024; Schmidt-Fajlik, 2023; Barrot, 2023). ChatGPT is brilliant at providing effective WCF because it is trained on a huge corpus of authentic texts, uses sophisticated in-built algorithms to discern sentence structure, is good at pattern recognition, and is programmed to self-improve.

RQ 2: How do ChatGPT and human feedback compare on the accuracy of the feedback?

The study sought to determine whether ChatGPT or humans are more accurate in identifying errors. In the comparative analysis, human-derived feedback outperformed ChatGPT's feedback by a narrow margin. The human rater identified more surface errors vis-à-vis ChatGPT. Moreover, the human rater missed fewer errors than ChatGPT. However, ChatGPT was close to the human rater on error-flagging accuracy. Thus, both the human and ChatGPT performed strongly in terms of accuracy. These findings resonate with a number of research inquiries into the error-flagging accuracy of ChatGPT and humans (Wu et al., 2023; Ranalli & Yamashita, 2022). Like the past studies referred to above, this study also establishes that ChatGPT is almost on par with humans in error-flagging accuracy.

Here, it would be in order to address a potential question that might arise when the findings flowing from addressing RQ1 and RQ2 are seen in relation to each other: Why did ChatGPT feedback prove to be more effective than human feedback in improving the accuracy of students' writing when humans outperformed ChatGPT in error flagging, albeit by a narrow margin? This apparent conflict between the results pertaining to RQ1 and RQ2 can be explained in terms of the following two factors: Firstly, the modality of ChatGPT feedback differed from human feedback in important ways. ChatGPT's corrections were delivered instantly in an interactive manner, allowing students to engage in a dialogic process not available with human feedback. In their dialogues with ChatGPT, the participants had the opportunity to ask supplementary questions arising out of ChatGPT WCF and receive answers in real time. This immediacy and interactivity would have helped participants deal effectively with the errors they had made. Secondly, it was noted that on most occasions, ChatGPT feedback was more comprehensive and illustrative than human feedback. Such feedback, combining detailed explanations with numerous examples, afforded participants a clearer understanding of the linguistic problems in their writing and solid solutions to those problems. This lucid awareness of both the problems and the solutions eventually led the students to make fewer errors as they composed in English.

RQ 3: How do the Saudi EFL learners feel about the efficacy of ChatGPT as a tool for Written Corrective Feedback?

The thematic analysis of the transcripts of semi-structured interviews conducted with six students from the experimental group revealed that overall, the Saudi students carried a positive view of the affordances of ChatGPT. In fact, most of the interviewees were deeply impressed by ChatGPT's accuracy of error identification and the clarity of error explanation. These students considered ChatGPT on par with human feedback in terms of error-flagging accuracy. One of the interviewees, however, did not think this highly about the effectiveness of ChatGPT. His perception was that while ChatGPT was really good at correctly identifying errors in grammar, lexis, and punctuation, human feedback was better than that of ChatGPT in terms of the clear explanation of errors. The findings outlined above are in line with many other studies on the subject of student perceptions of the efficacy of ChatGPT as a feedback tool (Steiss et al., 2024; Shi & Aryadoust, 2023).

6. Recommendations of the Study

Based on the findings and conclusions of this study, the following recommendations can be made for the stakeholders. First, since ChatGPT established its affordances in helping the participants write with greater linguistic accuracy, educators may consider integrating ChatGPT into their academic writing instruction scheme. They may adopt it as a co-feedback provider on the learners' English writing, allowing the students to become aware of their surface errors instantly. This study recommends ChatGPT as a co-feedback provider rather than the sole feedback provider for two reasons. One, in nuanced contexts, for instance the deliberate, stylistic use of sentence fragments, the human feedback provider, rather than ChatGPT, will catch the context and provide the appropriate feedback; and, two, humans have been found to be better feedback providers on non-surface errors. Second, the future versions of AWCF tools like ChatGPT should have an option that would let the learner specify what types of surface errors he or she wishes to concentrate on or get thorough elucidation for. This will make the AWCF more robust and in tune with the learning needs of specific students. Finally, if a decision is made to bring AWCF to writing classrooms, the students as well as the instructors should be trained to use ChatGPT or a similar AWCF tool for writing. The learners and instructors should be given training sessions, familiarizing them with features and powers of digital AWCF tools. This way the stakeholders will be better equipped to leverage the strengths and understand the limitations of AI-powered AWCF tools.

7. Limitations of the Study

This study has a few limitations that must be noted. Firstly, the findings are based on a relatively small sample size, which restricts the generalizability of the results. Secondly, while both groups engaged in the same number of writing sessions (two per week, each lasting 1.5 hours) and received the same frequency of feedback (once per session), the nature of the feedback differed. ChatGPT was able to provide immediate and interactive corrective feedback during the sessions, whereas human evaluators could not always deliver feedback with the same speed or level of elaboration. This difference in modality, rather than frequency, may have influenced the results.

A third limitation concerns perception bias. Some interviewees may have overstated ChatGPT's affordances, assuming that because the chatbot is AI-powered, it must be efficient. A further limitation concerns the interrater reliability procedure. Reliability was assessed using a subset of 12 essays, representing approximately 23% of the total dataset. This sample was selected to provide a manageable yet balanced cross-section of student work across both groups, allowing for a consistency check without re-evaluating the entire corpus. While this approach offered useful evidence of consistency between ChatGPT and human evaluators, future research could strengthen reliability by applying interrater checks to a larger proportion of the data or to the full dataset.

Finally, it can be assumed that the quality of human corrective and explanatory feedback varied among evaluators, and this variability may have influenced the results. Differences in scholarly standards and levels of competence across human evaluators could have introduced inconsistencies that affected the comparison with ChatGPT's feedback. Future research could investigate these sources of variability more systematically.

8. Conclusion

The aim of this investigation was to examine how effective ChatGPT's written corrective feedback was compared with the human written corrective feedback in enhancing the linguistic accuracy of writing produced by Saudi EFL learners. The quantitative findings showed a statistically significant improvement in the linguistic accuracy of writings produced by the participants who received ChatGPT's corrective feedback, as reflected in the post-test mean score of 66.15 for the experimental group, compared with posttest mean score of 56.48 for the control group. This statistical finding supports the view that ChatGPT can be harnessed as an effective technological supplement or alternative to human corrective feedback as Saudi EFL learners continue with their writing enhancement pursuits.

While the quantitative results indicated the relatively better impact of ChatGPT's corrective feedback on the participants' linguistic accuracy vis-à-vis the impact of human corrective feedback, the qualitative results revealed the strengths and weaknesses of both feedback sources. Here, ChatGPT not only established itself as a corrective feedback source almost on par with human corrective feedback in terms of accurate error-flagging capability but also garnered appreciation for its error-flagging and fixing capabilities from most of the interviewees. The interviewees praised ChatGPT for its real-time, mostly-accurate feedback which often fostered a richer comprehension of their surface errors. Many interviewees remarked that the error explanations provided by ChatGPT, though occasionally less clear than the explanatory feedback provided by humans, overall made valuable contribution to their written English accuracy.

As for the strength of human corrective feedback, it outperformed ChatGPT's feedback with a narrow margin in terms of error-flagging accuracy. The findings of this study suggest that ChatGPT was a more effective source of feedback compared with human corrective feedback. Such a move will enhance the written English of Saudi EFL students.

At the end, a cautionary note would be in order: while ChatGPT carries the same level of expertise across contexts, human evaluators' expertise would naturally vary across contexts. Therefore, any conclusions about feedback effectiveness flowing from a comparison of ChatGPT and humans should be viewed with a little caution.

Acknowledgments

This work was funded by the University of Jeddah, Jeddah City, Saudi Arabia, under grant No. (UJ-24-SHR-3333-1). The author acknowledges thanks to the University of Jeddah's technical and financial support. He would like to extend his heartiest thanks to the instructors and students who willingly participated in the study.

Authors' contributions

Dr. Fawaz Al-Mahmud is the sole author of this research paper.

Funding

This work was funded by the University of Jeddah, Jeddah City, Saudi Arabia, under grant No. (UJ-24-SHR-3333-1).

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Sciedu Press.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

Artificial Intelligence Statement: The author used Grammarly only for language editing and confirms that no AI-generated content was included; full responsibility for the work rests with the author.

References

Anderson, J. A., & Ayaawan, A. E. (2023). Formative feedback in a writing programme at the University of Ghana. In *African perspectives on the teaching and learning of English in higher education* (pp. 197-213). *Routledge*. https://doi.org/10.4324/9781003279433-15

Barrot, J. S. (2023). Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy. *Computer Assisted Language Learning, 36*(4), 584-607. https://doi.org/10.1080/09588221.2021.1936071

Barrot, J. S. (2024). ChatGPT as a language learning tool: An emerging technology report. *Technology, Knowledge and Learning, 29*(2), 1151-1156. https://doi.org/10.1007/s10758-023-09711-4

Bitchener, J., & Ferris, D. R. (2012). *Written corrective feedback in second language acquisition and writing.* https://doi.org/10.4324/9780203832400

Butterfuss, R., Roscoe, R. D., Allen, L. K., McCarthy, K. S., & McNamara, D. S. (2022). Strategy uptake in writing pal: Adaptive feedback and instruction. *Journal of Educational Computing Research, 60*(3), 696-721. https://doi.org/10.1177/07356331211045304

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language testing, 32*(3), 385-405. https://doi.org/10.1177/0265532214565386

Creswell, J. W. (2018). Educational research: Planning, conducting, and evaluating quantitative and qualitative research. Pearson.

Dizon, G., & Gayed, J. M. (2021). Examining the impact of Grammarly on the quality of mobile l2 writing. *jalt call Journal, 17*(2), 74-92. https://doi.org/10.29140/jaltcall.v17n2.336

Fokides, E., & Peristeraki, E. (2024). Comparing ChatGPT's correction and feedback comments with that of educators in the context of primary students' short essays written in English and Greek. *Education and Information Technologies*, 1-45. https://doi.org/10.1007/s10639-024-12912-8

Ghaemi, H., & Bayati, M. (2021). Software technology and writing skills improvement of intermediate EFL learners. *Journal of Research in Techno-based Language Education, 1*(2), 15-28.

Graham, S. (2019). Changing how writing is taught. *Review of Research in Education, 43*(1), 277-303. https://doi.org/10.3102/0091732X18821125

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The elementary school journal, 115*(4), 523-547. https://doi.org/10.1086/681947

Han, T., & Sari, E. (2024). An investigation on the use of automated feedback in Turkish EFL students' writing classes. *Computer Assisted Language Learning, 37*(4), 961-985. https://doi.org/10.1080/09588221.2022.2067179

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge grammar of the English language*. Cambridge University Press. https://doi.org/10.1017/9781316423530

Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies, 28*(12), 15873-15892. https://doi.org/10.1007/s10639-023-11834-1

Liao, H. C. (2016). Using automated writing evaluation to reduce grammar errors in writing. *Elt Journal, 70*(3), 308-319. https://doi.org/10.1093/elt/ccv058

Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: how may AI and GPT impact academia and libraries*? Library hi tech news, 40*(3), 26-29. https://doi.org/10.1108/LHTN-01-2023-0009

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

Mizumoto, A., Shintani, N., Sasaki, M., & Teng, M. F. (2024). Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment. *Research Methods in Applied Linguistics, 3*(2), 100116. https://doi.org/10.1016/j.rmal.2024.100116

Panadero, E., Alqassab, M., Fernández Ruiz, J., & Ocampo, J. C. (2023). A systematic review on peer assessment: intrapersonal and interpersonal factors. *Assessment & Evaluation in Higher Education, 48*(8), 1053-1075. https://doi.org/10.1080/02602938.2023.2164884

Peterson, M., & Jabbari, N. (2024). Contemporary computer assisted language learning: A view from the frontiers of research. In Frontiers in Technology-Mediated Language Learning (pp. 1-6). *Routledge*. https://doi.org/10.4324/9781003395218-1

Pfau, A., Polio, C., & Xu, Y. (2023). Exploring the potential of ChatGPT in assessing L2 writing accuracy for research purposes. *Research Methods in Applied Linguistics, 2*(3), 100083. https://doi.org/10.1016/j.rmal.2023.100083

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Ranalli, J., & Yamashita, T. (2022). *Automated written corrective feedback: Error-correction performance and timing of delivery.* https://doi.org/10.64152/10125/73465

Saldaña, J. (2021). Coding techniques for quantitative and mixed data. *The Routledge reviewer's guide to mixed methods analysis*, 151-160. https://doi.org/10.4324/9780203729434-14

Schmidt-Fajlik, R. (2023). ChatGPT as a grammar checker for Japanese English language learners: A comparison with Grammarly and ProWritingAid. *AsiaCALL Online Journal, 14*(1), 105-119. https://doi.org/10.54855/acoj.231417

Shi, H., & Aryadoust, V. (2023). *A systematic review of automated writing evaluation*.

Soltanpour, F., & Valizadeh, M. (2018). Revision-mediated and attention-mediated feedback: Effects on EFL learners' written syntactic accuracy. *Advances in Language and Literary Studies, 9*(4), 83-91. https://doi.org/10.7575/aiac.alls.v.9n.4p.83

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology, 14*, 1260843. https://doi.org/10.3389/fpsyg.2023.1260843

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

Thi, N. K., & Nikolov, M. (2022). How teacher and Grammarly feedback complement one another in Myanmar EFL students' writing. *The Asia-Pacific Education Researcher, 31*(6), 767-779. https://doi.org/10.1007/s40299-021-00625-2

Wang, S. (2011, December). On English Writing Feedback Provided by Writing Roadmap™ 2.0 Automated Evaluation System in EFL Classroom. In 2011 IEEE Ninth International Conference on Dependable, *Autonomic and Secure Computing* (pp. 1179-1183). IEEE. https://doi.org/10.1109/DASC.2011.191

Wang, W., & Li, S. (2021). Corrective feedback and learner uptake in American ESL and Chinese EFL classrooms: A comparative study. *Language, Culture and Curriculum, 34*(1), 35-50. https://doi.org/10.1080/07908318.2020.1767124

Wilson, J., & MacArthur, C. (2024). Exploring the role of automated writing evaluation as a formative assessment tool supporting self-regulated learning in writing. *In The Routledge international handbook of automated essay evaluation* (pp. 197-220). Routledge. https://doi.org/10.4324/9781003397618-14

Wu, H., Wang, W., Wan, Y., Jiao, W., & Lyu, M. (2023). Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark. *arXiv preprint arXiv:2303.13648*.

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