ASIA PACIFIC JOURNAL
OF EDUCATORS AND EDUCATION

Volume 41, Issue 1, 2026
https://doi.org/10.21315/apjee2026.41.1.8

Research Article:

Human vs. Machine Feedback: Evaluating ChatGPT-4 in the Assessment of Secondary EFL Learners' Writing

Yasemin Gok Acan[1] and Aysegul Liman Kaban[2]*
[1]Computer Education and Instructional Technologies Department, Bahcesehir University, 34349 Istanbul, Türkiye
[2]STEM Education Department, Mary Immaculate College, V94 VN26 Limerick, Ireland

*Corresponding author: aysegul.limankaban@mic.ul.ie

ABSTRACT

This mixed-methods study compared ChatGPT-4 generated feedback with teacher feedback in assessing secondary school EFL learners' writing. Conducted in a state secondary school in Istanbul, Türkiye, the study involved 53 5th-grade students who completed weekly writing tasks over six weeks. One class received teacher feedback, while two classes received feedback from ChatGPT-4. Students revised their texts based on the feedback, and changes in writing performance were analysed quantitatively. In addition, semi-structured interviews with 45 students were conducted to explore their feedback preferences and perceptions. The quantitative findings showed that both feedback types supported improvement in students' writing, although teacher feedback produced more consistent gains across the six-week period. A strong positive correlation was found between ChatGPT-4 and teacher scores in both pre- and post-feedback assessments, suggesting that AI-generated scoring aligned closely with human evaluation. However, the qualitative findings revealed that students generally preferred teacher feedback, describing it as more personal, motivating, and easier to trust. ChatGPT-4 feedback was appreciated for its speed, clarity, and accessibility, but was also seen as less detailed and less emotionally engaging. The findings suggest that generative AI can serve as a useful formative feedback tool in EFL writing instruction, but it does not replace the pedagogical and relational strengths of teacher feedback. A hybrid model that combines the efficiency of AI with the contextual sensitivity of human feedback may offer the most effective approach for supporting writing development in secondary language classrooms.

Keywords: AI-generated feedback, ChatGPT-4, EFL writing instruction, formative assessment, human–machine collaboration

Published: 9 June 2026

To cite this article: Gok Acan, Y., & Liman Kaban, A. (2026). Human vs. machine feedback: Evaluating ChatGPT-4 in the assessment of secondary EFL Learners' writing. *Asia Pacific Journal of Educators and Education, 41*(1), 159–187. https://doi.org/10.21315/apjee2026.41.1.8

© Penerbit Universiti Sains Malaysia, 2026. This work is licensed under the terms of the Creative Commons Attribution (CC BY) (http://creativecommons.org/licenses/by/4.0/).

INTRODUCTION

Feedback plays a critical role in shaping students' learning, particularly in writing, where formative guidance supports the iterative development of language, structure, and ideas. In the context of English as a Foreign Language (EFL) education, effective feedback is instrumental in helping learners improve both linguistic accuracy and rhetorical sophistication (Hattie & Timperley, 2007; Guo & Wang, 2024). Traditionally, teachers have been the primary providers of such feedback, valued for their ability to offer personalised, context-sensitive, and affectively attuned responses (Farrokhnia et al., 2026). However, rising student-to-teacher ratios, time constraints, and increasing workloads especially in resource-constrained or large-classroom environments have made timely, individualised feedback increasingly difficult to deliver at scale (Guo et al., 2024). In response to these challenges, technological tools, particularly those powered by artificial intelligence (AI), are gaining attention as potential feedback providers. Automated writing evaluation (AWE) systems, including tools like Grammarly and Pigai, have been in use for over a decade, offering instant feedback on grammar, coherence, and structure. Yet these systems have been criticised for focusing primarily on surface-level features, providing generic or overly prescriptive comments, and failing to address deeper elements such as argument quality or rhetorical intent (Shi & Aryadoust, 2024; Banihashem et al., 2024).

The emergence of large language models (LLMs) such as ChatGPT-4 represents a significant evolution in this space. Unlike earlier AWE systems, ChatGPT can generate human-like feedback with greater fluency, contextuality, and responsiveness to diverse prompts and genres. Early studies suggest that generative AI can provide feedback that aligns reasonably well with human evaluators on surface-level aspects of writing (Steiss et al., 2024), and students often appreciate the speed and accessibility of AI-generated responses (Escalante et al., 2023). Nevertheless, concerns persist regarding the diagnostic depth, pedagogical validity, and emotional resonance of AI-generated feedback, particularly in settings where language learning is deeply interpersonal and culturally situated (Guo & Wang, 2024).

Despite the growing discourse around AI in education, empirical research comparing ChatGPT-generated feedback to that of experienced EFL teachers—particularly at the secondary level—remains limited. Most existing studies focus on university-level learners or programming and technical education, leaving a gap in our understanding of how generative AI can support writing development among younger, language-learning students in school contexts (Banihashem et al., 2024; Escalante et al., 2023). The integration of generative AI tools such as ChatGPT into educational settings is no longer hypothetical, and it is actively shaping classroom practices. While much of the existing research focuses on university-level learners, this study contributes a grounded investigation into how AI-generated feedback is received and utilised by secondary school students in EFL contexts. This focus on younger learners fills a notable gap in the literature and offers insights that can inform school-level policy and instructional design.

This study aims to address this gap by investigating how feedback generated by ChatGPT-4 compares to traditional teacher feedback in the context of secondary EFL writing. Specifically, the study examines the impact of both feedback types on students' writing performance and explores students' perceptions of their usefulness and credibility. By situating these findings within the current literature, this research offers practical insights into the pedagogical potential and limitations of human–machine feedback collaboration in language education.

The research questions are as follows:

> RQ1: How does GenAI feedback compare to teacher feedback in improving the writing proficiency of EFL students in secondary school in terms of content, organisation, and language?
>
> a) Is there a statistically significant difference between the pre-feedback and post-feedback scores of students when they get feedback from GenAI?
> b) Is there a statistically significant difference between the pre-feedback and post-feedback scores of students when they get feedback from a teacher?
>
> RQ2: What is the level of correlation between GenAI and teacher scores in assessing EFL students' writings in terms of content, language, and organisation?
>
> RQ3: What are secondary school EFL students' preferences between GenAI and teacher feedback, and what reasons do they provide for their choices?

LITERATURE REVIEW

The Importance of Written Corrective Feedback

Written corrective feedback (WCF) has long occupied a central place in second language writing research and continues to be widely discussed in relation to how learners improve the accuracy and quality of their written production. Crosthwaite et al. (2022) define WCF as handwritten or electronic markings used to address linguistic errors across different levels of writing. Although its pedagogical value has been debated, WCF remains one of the most established forms of instructional support in L2 writing classrooms. A major point of contention emerged with Truscott's (1996) claim that corrective feedback was ineffective and potentially harmful to language development. However, this position has been strongly challenged by subsequent empirical research, which has generally demonstrated that WCF can make an important contribution to second language writing development when it is provided in purposeful and pedagogically appropriate ways.

In particular, later studies have reinforced the view that WCF plays a valuable role in supporting learners' grammatical development. Bitchener and Ferris (2012), for example,

argue that written corrective feedback remains a key component of second language writing instruction because it helps learners notice, understand, and address language-related problems in their writing. Supporting this view, Bitchener (2008) found that direct written corrective feedback led to significant improvement in learners' use of English articles, with these gains sustained over a two-month period. This suggests that well-targeted feedback can have both immediate and lasting effects. Along similar lines, Rassaei and Jabbarpoor (2025) reported that explicit corrective feedback was more effective than recasts in supporting Persian EFL learners, pointing to the importance of visibility and clarity in the feedback process. Taken together, these studies indicate that feedback is particularly effective when learners can clearly identify the problem and understand the intended correction.

At the same time, the effectiveness of WCF cannot be understood independently of how learners engage with it. Research has shown that student responses to feedback are shaped by factors such as language proficiency and writing genre. Kang and Han (2015) and Zheng et al. (2023) found that lower-proficiency learners often engaged with teacher feedback behaviourally and affectively, yet faced difficulties at the cognitive level because of their more limited linguistic resources. In other words, students may be willing to respond to feedback and may even value it but still struggle to process it deeply enough to revise effectively. Building on this line of work, Wang et al. (2025) emphasise that feedback should be designed in ways that support learners' affective, behavioural, and cognitive engagement. They also suggest that alternative or complementary forms of feedback, including peer and automated feedback, may help address these different dimensions of engagement more effectively. This broader perspective is especially relevant in current discussions of AI-supported feedback, as it highlights that the impact of feedback depends not only on its accuracy but also on how accessible, interpretable, and usable it is for learners.

The Emergence of ChatGPT as a Tool for Language Education

ChatGPT, developed by OpenAI, has rapidly established itself as one of the most consequential developments in educational technology, with its role in language learning evolving substantially across successive model generations from GPT-1 to GPT-4 (Teng, 2024). Unlike earlier automated writing evaluation (AWE) systems that operated primarily at the surface level of grammar and syntax, large language models (LLMs) such as ChatGPT are capable of generating nuanced, contextually responsive, and multi-dimensional feedback, representing what many scholars describe as a paradigm shift in AI-assisted language education (Shi & Aryadoust, 2024; Xiao et al., 2026). Survey evidence confirms that students are already engaging extensively with the tool, particularly for writing-related tasks (von Garrel & Mayer, 2023), and a growing body of empirical work attests to its value across writing support, vocabulary development, and formative assessment (Kohnke et al., 2023; Kostka & Toncelli, 2023).

A principal driver of this uptake is ChatGPT's capacity for personalised, on-demand feedback, a feature that addresses longstanding structural limitations in large EFL

classrooms where high student-to-teacher ratios make timely, individualised feedback difficult to deliver consistently (Guo & Wang, 2024). Unlike static AWE tools, ChatGPT can respond to diverse prompts and genre-specific writing tasks, producing feedback that addresses content, organisation, and language in an integrated way (Baskara, 2023; Dai et al., 2023). This flexibility has been linked to improvements in students' revision processes, motivation, and emotional engagement with writing tasks (Teng, 2024). Research further shows that learners who interact with AI-generated feedback demonstrate gains in learner autonomy and self-regulated writing behaviours, particularly when structured prompting strategies are incorporated into the instructional design (Agustini, 2023; Su et al., 2023).

Comparative research has substantially strengthened the evidence base for ChatGPT's educational value. Although human feedback continues to outperform AI across most qualitative dimensions of formative assessment, ChatGPT has been shown to perform competitively during the early drafting stages of writing development (Steiss et al., 2024), and to generate feedback comparable in consistency and detail to that of trained evaluators when guided by explicit rubric-based prompts. Studies by Yamashita (2025) and Mizumoto et al. (2024) reported strong correlations between AI- and human-generated scores, positioning ChatGPT as a credible tool for automated essay scoring (AES) and formative support at scale. Importantly, Zou et al. (2025) found that learners showed higher engagement with organisation-focused feedback when it came from ChatGPT, suggesting that the two feedback sources may have complementary rather than competing pedagogical strengths.

At the same time, the literature is candid about ChatGPT's limitations. A particularly significant concern is the phenomenon of AI hallucination, the generation of plausible sounding but factually inaccurate or contextually inappropriate feedback, which has been documented across multiple writing feedback studies (Yoon et al., 2023; Xiao et al., 2026). ChatGPT may also produce generic or formulaic responses that fail to account for the rhetorical complexity, cultural context, or genre-specific conventions of a given writing task (Evmenova et al., 2024; Pfau et al., 2023). These limitations are compounded by the risk of uncritical feedback uptake: when learners accept AI-generated suggestions without evaluative scrutiny, passive revision replaces the metacognitive engagement that is central to writing development (Zhan & Yan, 2025). Related to this, Barrot (2023) cautions that the convenience of AI feedback may reduce students' cognitive investment in the writing process, with implications for critical thinking, creative risk-taking, and the development of an authentic writerly voice.

For these reasons, current research increasingly advocates for a model of human-AI collaboration in EFL writing instruction, rather than the substitution of one feedback source for another. Su et al. (2023) propose that ChatGPT functions most effectively as a collaborative partner in the drafting and revision cycle, complementing rather than replacing teacher-mediated feedback on higher-order concerns. This position is reinforced by Steiss et al. (2024), who found that while ChatGPT's overall feedback quality approaches that of trained human evaluators in certain respects, teachers remain better

positioned to identify the most pedagogically noticeable issues and provide the kind of scaffolded instructional direction that promotes durable writing development. Realising the potential of such collaboration, however, requires deliberate attention to AI literacy, including students' capacity to critically evaluate AI outputs, recognise hallucinations, and make ethical decisions about which suggestions to adopt alongside transparent classroom policies, structured prompting instruction, and ongoing teacher oversight (Alzahrani, 2026; Gunes & Liman Kaban, 2025). Integrated thoughtfully, ChatGPT holds genuine promise as a tool for expanding access to timely, personalised formative feedback, provided its use is secured in sound pedagogical judgement and a sustained commitment to developing learners as critical, self-regulating writers.

Ethics of Using ChatGPT in EFL Writing

The integration of generative AI tools such as ChatGPT into EFL writing instruction raises a range of ethical concerns that demand careful pedagogical attention. Leading among these are questions of academic integrity, uncritical dependence, and the erosion of learner agency. Researchers including Casal and Kessler (2023), Chan (2023), Cotton et al. (2024), and Gunes and Liman Kaban (2025) have cautioned that the undisclosed or unreflective use of AI-generated content may undermine academic honesty, particularly when students submit AI-assisted work without independent cognitive investment. In response, institutions have increasingly introduced guidelines, restrictions, or outright bans on AI use in assessed writing (Gunes & Liman Kaban, 2025; Sharples, 2025; Teng, 2024). However, the emerging consensus in the literature suggests that prohibition alone is neither sufficient nor pedagogically sound. Rather, scholars advocate for authentically designed assessment tasks including reflective writing, visual interpretation, and experience-based narratives that are less susceptible to AI automation and more likely to elicit genuine student voice (Moorhouse et al., 2023; Rudolph et al., 2023).

A deeper and increasingly urgent concern relates to the cognitive consequences of AI dependency in writing contexts. When students routinely delegate drafting, structuring, or revising to ChatGPT, they risk bypassing the very cognitive processes that writing instruction is designed to cultivate. This phenomenon is commonly theorised as cognitive offloading, the externalisation of cognitive tasks onto technological tools, which reduces the mental effort learners invest in meaningful meaning-making (Gerlich, 2025). While cognitive offloading can serve legitimate scaffolding functions, its habitual use in writing contexts risks producing what Kosmyna et al. (2025) have termed cognitive debt: the accumulation of long-term cognitive costs, including diminished critical thinking, reduced creativity, and weakened authorial ownership that result from sustained over-reliance on AI-generated output. In their neurophysiological study, Kosmyna et al. (2025) found that participants who wrote essays using AI assistance exhibited up to 55% reduced neural connectivity compared to those who wrote independently, and 83% were unable to recall content from essays they had just produced with AI support. A related concept, cognitive surrender, captures the moment at which learners cease to critically evaluate AI-generated suggestions and instead accept them passively, forfeiting the evaluative judgement that

underpins genuine learning (Alzahrani, 2026; Shaw & Nave, 2026). These concerns are compounded by Gerlich's (2025) finding of a significant negative correlation between frequent AI tool use and critical thinking scores across 666 participants, with cognitive offloading identified as the primary mediating mechanism.

Barrot (2023) further argues that the fluency and accessibility of tools like ChatGPT can lead to shallow engagement with writing tasks, weakening students' capacity for independent critical analysis and creative expression. Cotton et al. (2024) caution that sustained dependence may gradually erode academic independence and learners' sense of authorship, while Ting et al. (2025) draw attention to how the prescriptive discursive stance adopted by AI feedback systems can inadvertently constrain learner agency and position the AI as an unquestioned epistemic authority. Taken together, these perspectives suggest that the ethical risks of generative AI in writing instruction extend well beyond questions of plagiarism detection, reaching into the architecture of cognition and the formation of academic identity.

To navigate these challenges, scholars increasingly emphasise the development of AI literacy as a foundational ethical and pedagogical response. AI literacy in this context includes not only technical proficiency but also the capacity for critical, evaluative engagement with AI-generated output, enabling learners to question, verify, and selectively incorporate AI suggestions rather than accept them as authoritative (Alzahrani, 2026; Gunes & Liman Kaban, 2025). Process-oriented assessment approaches, including annotated drafting sequences, oral reflective tasks, and collaborative writing, offer productive means of making students' cognitive processes visible and thereby sustaining ethical engagement (Casal & Kessler, 2023; Chan, 2023). Ultimately, the responsible integration of ChatGPT into EFL writing pedagogy requires a balanced approach in which the tool's acknowledged efficiency gains are accompanied by structured opportunities for critical reflection, independent reasoning, and the cultivation of writerly voice, ensuring that AI augments rather than replaces the deeper intellectual work that effective writing instruction demands.

Theoretical Framework

This study is guided by a feedback-oriented theoretical framework drawing on two complementary perspectives. First, Hattie and Timperley's (2007) model conceptualises feedback as information that helps learners understand their current performance, identify learning goals, and determine next steps for improvement. This perspective is particularly relevant to EFL writing, where effective feedback should not only point out errors but also support revision and language development. Second, the study is informed by feedback literacy research, which emphasises that feedback becomes pedagogically meaningful only when learners can interpret, value, and act on it. From this perspective, the effectiveness of feedback is shaped not only by its technical accuracy but also by its clarity, accessibility, relational quality, and perceived usefulness. Together, these perspectives provide a suitable framework for comparing teacher and AI-generated feedback, as they allow the study to examine both performance outcomes and students' engagement with feedback as part of AI-supported language learning.

METHODS

In this research, a mixed-method design was preferred to provide more comprehensive insight into the topic. In a mixed-method research design, both quantitative and qualitative techniques are applied within a single study (Johnson & Onwuegbuzie, 2004). Researchers using mixed-methods combine qualitative and quantitative approaches to explore both convergent and divergent findings (Creswell & Plano Clark, 2011). This study adopted a quasi-experimental design to investigate the effectiveness of ChatGPT feedback versus teacher feedback on secondary EFL students' writing performance. Quasi-experimental research lacks random assignment but attempts to match the control group closely to the treatment group (Lam & Wolfe, 2023). Due to practical constraints within the school, such as teacher availability and scheduling, group assignment was not randomised. One class received teacher feedback, while two classes received AI-generated feedback. To address potential imbalance, a baseline writing proficiency task was administered before the intervention. Results indicated comparable performance across the groups, and this was considered during the analysis to guarantee fair comparison. While convenience sampling was used, the implications of this design choice are discussed in the limitations section.

Convenience sampling was used, as three 5th-grade classes were readily accessible. One class received teacher feedback (control group), and two classes received ChatGPT feedback (experimental group). Pre- and post-feedback writing scores were collected for six weeks to evaluate improvement. To complement the quantitative data, qualitative data were collected through semi-structured interviews exploring student feedback preferences and perceptions. The quantitative and qualitative data were collected and analysed separately and later triangulated to obtain a more holistic understanding of the findings.

Setting

The research took place in a state secondary school in Istanbul, Türkiye, during March and April of the 2023–2024 academic year. Students had five English lessons per week. While students lacked personal computers, each classroom had a smartboard, and English instruction followed a national curriculum.

Participants

Fifty-three 5th-grade students from three classrooms (5A, 5B, 5C) participated in the study. Their average age was 10, and English proficiency levels were estimated between A1 and A2. Gender was mixed across classrooms. All three classes were taught by the same experienced female teacher. One class served as the control group, while the other two were experimental groups. The demographic distribution of participants and assessor information are provided in Tables 1 and 2.

Table 1. Demographic information of the participants

Class	Gender	No. of participants	Percentage (%)	Age
5A	Girls	8	47.06	10
	Boys	9	52.94	
5B	Girls	9	47.37	10
	Boys	10	52.63	
5C	Girls	8	47.06	10
	Boys	9	52.94	
	Girls	25	47.17	10
	Boys	28	52.83	

Table 2. Demographic and professional background of assessors

Assessors	Age	Years of experience	Levels of teaching
Teacher	30	7	K-12 levels
ChatGPT-4	*	*	*

Procedure

Figure 1 illustrates the step-by-step procedure for the research comparing ChatGPT-4 and teacher feedback in assessing the writing of secondary school EFL students. The study began with the selection of 53 fifth-grade students from three classrooms in a public school in Istanbul. These students were divided into two groups: the experimental group received feedback from ChatGPT-4, while the control group received traditional teacher feedback. Over six weeks, students completed weekly writing tasks related to the English curriculum. Since students lacked access to computers, their handwritten texts were typed by the teacher and then evaluated using a standardised rubric. Each group received feedback accordingly—either from the AI or the teacher and students revised their texts based on the feedback. Final drafts were again assessed to track improvement. To complement the quantitative data, semi-structured interviews were conducted with 45 students to explore their feedback preferences and reasoning. This multi-stage design allowed the researcher to systematically compare feedback efficacy and gain insights into student perceptions.

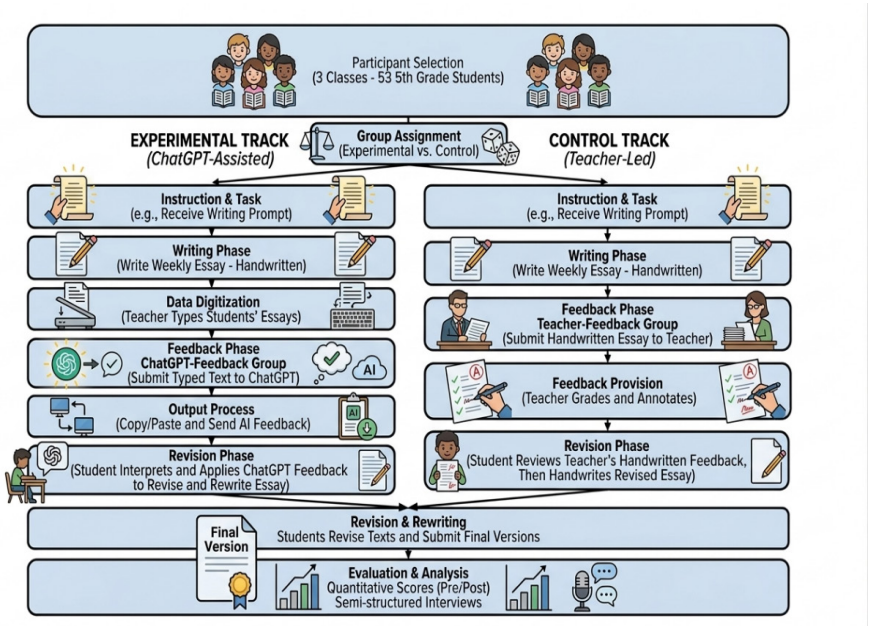

Figure 1. Procedure of the intervention (*Note:* The figure was created by using Gemini).

Data Collection Process

Students completed writing tasks based on their curriculum over six weeks. All work was handwritten, then typed by the teacher before being assessed by both ChatGPT-4 and the teacher using the same rubric. Feedback was returned to students' ChatGPT for the experimental group and the teacher for the control group and they revised their essays accordingly. This process was repeated weekly.

In May 2024, semi-structured interviews were conducted with 45 students who compared anonymised teacher and ChatGPT feedback on a final assignment. Students were asked to reflect on which type of feedback they preferred and why, with interviews recorded through note-taking.

Piloting

Pilot testing revealed that ChatGPT struggled to read handwritten texts and often produced linguistically complex feedback. To address this, handwritten essays were typed, and a customised prompt was developed to generate A1-level feedback in simple present

tense, incorporating emojis and motivational phrasing. The final prompt was tested and refined to ensure clarity and accessibility for learners.

Data Collection Instruments

1. ChatGPT-4: Chosen for its superior consistency and nuanced feedback compared to ChatGPT-3.5 (Bewersdorff et al., 2023; OpenAI, 2023).
2. Rubric: The Cambridge English KEY for Schools rubric was used to assess writing on content, organisation, and language (out of 5 points).
3. Prompting: A customised ChatGPT prompt emphasised simplicity, encouragement, and clarity.
4. Semi-Structured Interviews: Used to gather student reflections and preferences concerning feedback.

Data Analysis Process

Quantitative data were analysed using IBM SPSS. The Shapiro-Wilk test assessed data normality. For normally distributed data, paired samples t-tests were used; otherwise, Wilcoxon Signed Ranks tests were applied. Pearson correlation tested score alignment between ChatGPT and teacher assessments. Bar charts supported the correlation analysis. An inductive thematic analysis was conducted on the open-ended student survey responses. Two researchers independently coded the data, identifying recurring themes related to feedback preferences, perceived usefulness, and emotional responses. The coding process involved iterative refinement, and final themes were agreed upon collaboratively. Sample codes included "clarity of feedback", "personal connection", and "motivation". This approach ensured transparency and reliability in the qualitative findings.

Reliability and Validity

Reliability was supported through piloting, rubric standardisation, and dual scoring. Writing samples were assessed using the Cambridge KEY rubric, which is appropriate for the participants' proficiency level. To ensure scoring reliability, two independent raters evaluated the samples. A calibration session was conducted before scoring to align interpretations of the rubric. To ensure consistency in scoring, two independent raters evaluated all writing samples using a standardised rubric designed for EFL secondary learners. Before scoring, the raters participated in a calibration session to align their interpretations of the rubric criteria. Inter-rater reliability was assessed using Cohen's kappa, yielding a value of 0.82, which indicates strong agreement. Any discrepancies were discussed and resolved through consensus. Normality tests guided appropriate statistical test selection (see Tables 3–6).

Table 3. The normality test for RQ1a

Intervention	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
w1pre	.170	31	.022	.934	31	.056
w1post	.122	31	.200*	.942	31	.095
w2pre	.237	30	.000	.917	30	.023
w2post	.217	30	.001	.907	30	.013
w3pre	.219	32	.000	.893	32	.004
w4pre	.258	26	.000	.847	26	.001
w4post	.257	26	.000	.793	26	.000
w5pre	.154	30	.066	.914	30	.018
w5post	.256	30	.000	.866	30	.001
w6pre	.164	33	.025	.909	33	.009
w6post	.157	33	.037	.928	33	.030

Note: a = Lilliefors significance correction

Similarly, Table 4 shows the normality test to examine the students' pre-feedback and post-feedback writing scores given by the teacher. According to the normality test, the first and the sixth week data were normally distributed while the remaining weeks were not.

Table 4. The normality test for RQ1b

Intervention	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
w1pre	.149	16	.200*	.918	16	.155
w1post	.198	16	.094	.919	16	.163
w2pre	.258	19	.002	.826	19	.003
w2post	.355	19	.000	.680	19	.000
w3pre	.281	16	.001	.789	16	.002
w3post	.302	16	.000	.635	16	.000
w4pre	.252	19	.003	.785	19	.001
w4post	.281	19	.000	.803	19	.001
w5pre	.275	14	.005	.758	14	.002
w5post	.298	14	.001	.710	14	.000
w6pre	.177	18	.200*	.933	10	.480
w6post	.140	18	.200*	.942	18	.317

Note: a = Lilliefors significance correction; * = this is a lower bound of the true significance.

Table 5 shows the normality test for the data consisting of all the pre-feedback scores provided by both GenAI and the teacher. According to this test, the data was normally distributed, and thus the Pearson Correlation test was conducted to see the correlation between the data sets.

Table 5. The test of normality for RQ2 (pre-feedback scores)

Variable	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
GPT-pre(total)	.119	318	.000	.969	318	.000
Teacher-pre(total)	.119	318	.000	.950	318	.000

Note: Lilliefors Significance Correction

Table 6 shows the normality test for the data consisting of all the post-feedback scores provided by both GenAI and the teacher. According to this test, the data was normally distributed, which led the Pearson Correlation test to see the correlation between the two sets of scores.

Table 6. The test of normality for RQ2 (post-feedback scores)

Variable	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
GPT-post(total)	.145	318	.000	.907	318	.000
Teacher-post(total)	.168	318	.000	.867	318	.000

Note: Lilliefors Significance Correction

Ethical Considerations

Ethical approval for the study was obtained from the Bahcesehir University Research Ethics Committee. Informed consent was secured from all participants and their legal guardians. Students were informed of their right to withdraw at any time, and all data were anonymised to protect participant privacy. The study adhered to GDPR guidelines and institutional policies on research with minors.

FINDINGS

To enhance interpretability, effect sizes have been included alongside *p*-values for all statistically significant findings. For example, the improvement in writing scores in the AI-generated feedback group yielded a medium effect size (Cohen's d = 0.56), suggesting practical significance beyond statistical results. This addition helps readers better understand the magnitude of observed changes.

Findings of Research Question 1

Because the Week 1 data were normally distributed, a paired-samples t-test was conducted to compare students' pre-feedback and post-feedback scores following ChatGPT feedback (see Table 7). Although the mean post-feedback score was slightly higher than the mean pre-feedback score, the difference was not statistically significant, as the *p*-value (.136) was greater than .05. This indicates that students did not show a significant improvement in writing scores after receiving ChatGPT feedback in the first week.

Table 7. The paired samples test for RQ1

Pair		Paired differences					*t*	df	Sig. (2-tailed)
		Mean	S. D.	Std. error mean	95% confidence interval of the difference				
					Lower	Upper			
Pair 1	w1post–w1pre	.903	3.280	.589	–2.106	.300	–1.533	30	.136

The data was not normally distributed in the remaining weeks, so the Wilcoxon Signed Ranks Test was applied instead of the Paired Samples Test. As shown in Table 8, Wilcoxon tests revealed statistically significant improvements in Weeks 2, 3, 4, and 6 for the AI-generated feedback group ($p < .05$), indicating that ChatGPT feedback contributed to measurable gains in writing performance. However, weeks 1 and 5 did not show significant changes, suggesting variability in feedback impact.

Table 8. Wilcoxon signed ranks test for RQ1a

Test statistic	W2POST–W2PRE	W3POST–W3PRE	W4POST–W4PRE	W5POST–W5PRE	W6POST–W6PRE
z	–2.022[b]	–2.387[b]	–3.071[b]	–1.028[b]	–2.543[b]
Asymp. Sig. (2-tailed)	.043	.017	.002	.304	.011

Note: a = Wilcoxon signed ranks test; b = based on negative ranks.

To summarise, in the first and fifth weeks, students did not significantly improve their scores after receiving feedback from ChatGPT. However, in the second, third, fourth, and sixth weeks, students showed statistically significant improvements following ChatGPT feedback. Since the data were normally distributed in Weeks 1 and 6, a Paired Samples Test was applied to examine the differences between post-feedback and pre-feedback scores given by the teacher, as presented in Table 9. In week 1, the results indicate that post-feedback scores were higher than pre-feedback scores, and the *p*-value (0.010) was less than 0.05, indicating a statistically significant difference. This suggests that students

significantly improved their scores after receiving teacher feedback in the first week. In contrast, although post-feedback scores in week 6 were slightly higher than pre-feedback scores, the *p*-value (0.489) exceeded 0.05, indicating that the difference was not statistically significant. Therefore, students did not show a significant improvement in their scores after receiving teacher feedback in the sixth week.

Table 9. Paired samples test for RQ1b

Pair comparison		Paired differences					*t*	df	Sig. (2-tailed)
		Mean	S. D.	Std. error mean	95% confidence interval of the difference				
					Lower	Upper			
Pair 1	w1post–w1pre	2.187	2.994	.748	–3.783	–.592	–2.923	15	.010
Pair 2	w6post–w6pre	.667	4.000	.943	–2.656	1.322	–.707	17	.489

The data was not normally distributed in Weeks 2, 3, 4 and 5, and therefore the Wilcoxon Test was applied, which was presented in Table 10. Teacher's feedback showed consistent effectiveness in Weeks 3, 4, and 5, reinforcing its pedagogical value. The non-significant results in Weeks 2 and 6 may reflect contextual factors or student variability.

Table 10. Wilcoxon Signed Ranks Test for RQ1b

Test statistic	W2POST–W2PRE	W3POST–W3PRE	W4POST–W4PRE	W5POST–W5PRE
z	–1.274[b]	–2.442[b]	–2.279[b]	–2.282[b]
Asymp. sig. (2-tailed)	.203	.015	.023	.023

Notes: a = Wilcoxon signed-rank test; b = based on negative ranks

Overall, in Weeks 2 and 6, the students did not significantly increase their scores after getting feedback from the teacher; however, in Weeks 1, 3, 4, and 5, they increased their scores after getting feedback.

Table 11 presents the descriptive statistics for the pre-feedback scores given by both ChatGPT and the teacher. For the pre-feedback writing tasks, the mean of ChatGPT scores is 8.78, and the mean of teacher scores is 9.57, which shows that the teacher scores are a bit higher than ChatGPT scores. In the Pearson Correlation Test, the *p*-value is 0.000, showing that the correlation between the two sets of data was statistically significant, and *r* = 870 indicates a strong positive correlation. A correlation of 0.870 means that while one set of scores increases, the other also increases.

Table 11. Descriptive statistics for RQ2 (pre-feedback scores)

Variable	Mean	SD	*N*
GPT-pre(total)	8.78	2.872	318
Teacher-pre(total)	9.57	3.340	318

Table 12. Pearson correlation test for RQ2 (pre-feedback scores)

Variable		GPT-pre(total)	Teacher-pre(total)
GPT-pre(total)	Pearson correlation	1	.870**
	Sig. (2-tailed)		.000
	N	318	318
Teacher-pre(total)	Pearson correlation	.870**	1
	Sig. (2-tailed)	.000	
	N	318	318

Note: **Correlation is significant at the 0.01 level (2-tailed)

The second research question aims to find out the correlation between the teacher's scores and ChatGPT. To clarify this, the pre-feedback and post-feedback scores were tested independently. Both sets of data were normally distributed; therefore, the Pearson Correlation Test was preferred to apply, which was presented in Table 12. The sample size is 318 after excluding the students who had missing scores.

Post-feedback scores were also normally distributed. Table 13 presents the descriptive statistics of the post-feedback scores given by ChatGPT and the teacher. The mean of ChatGPT scores is 9.10 and the mean of teacher scores is 9.81, which shows that the teacher scores are a bit higher than ChatGPT scores. Table 14 presents the Pearson Correlation Test showing the correlation between the post-feedback scores given by the teacher and ChatGPT. According to the Pearson Correlation Test, $r = .858$ and $p = .000$ indicate that there is a strong positive correlation between the two sets of scores.

Table 13. descriptive statistics for RQ2 (post-feedback scores)

Variable	Mean	SD	*N*
GPT-post(total)	9.10	4.152	318
Teacher-post(total)	9.81	4.662	318

Table 14. Pearson correlation test for RQ2 (post-feedback scores)

Variable		GPT-post(total)	Teacher-post(total)
GPT-post(total)	Pearson correlation	1	.858**
	Sig. (2-tailed)		.000
	N	318	318

(Continued on next page)

Table 14. *(Continued)*

Variable		GPT-post(total)	Teacher-post(total)
Teacher-post(total)	Pearson correlation	.858	1
	Sig. (2-tailed)	.000	
	N	318	318

Note: **Correlation is significant at the .001 level (2-tailed)

Beside the correlation test, the following bar charts revealed the similarity and difference of the scores between the teacher and ChatGPT focusing on the criteria which were content, organisation, and language.

Figure 2 shows the scoring alignment between ChatGPT and the teacher for the content part in the evaluated writing tasks. ChatGPT and the teacher assigned identical scores for 317 tasks, indicating a strong level of agreement. However, the teacher gave scores that were 1 point higher than ChatGPT for 184 tasks and 2 points higher for 24 tasks. Conversely, ChatGPT assigned scores that were 1 point higher than the teacher for 69 tasks and 2 points higher for 1 task. This highlights some differences in scoring tendencies between ChatGPT and the teacher.

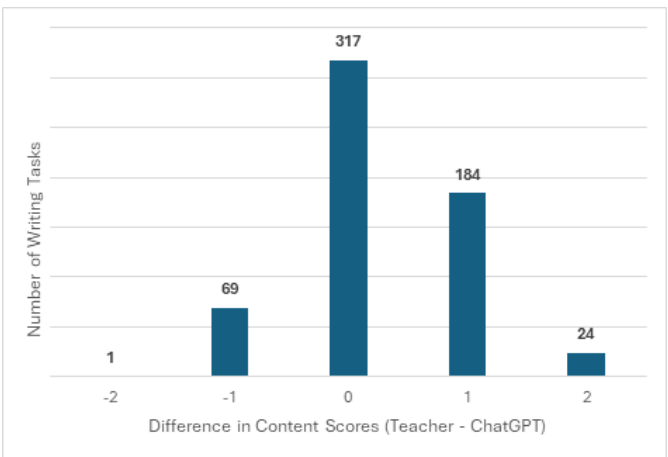

Figure 2. Difference in content scores (Teacher – ChatGPT)

Figure 3 shows the scoring alignment between ChatGPT and the teacher for the organisation part. It is seen that for 281 tasks, ChatGPT and the teacher gave identical scores, highlighting a strong. In 188 tasks, the teacher scored 1 point higher than ChatGPT, while in 32 tasks, the teacher scored 2 points higher. Additionally, for 3 tasks, the teacher assigned scores that were 3 points higher than ChatGPT. Conversely, ChatGPT scored 1 point higher than the teacher in 86 tasks and 2 points higher in 5 tasks.

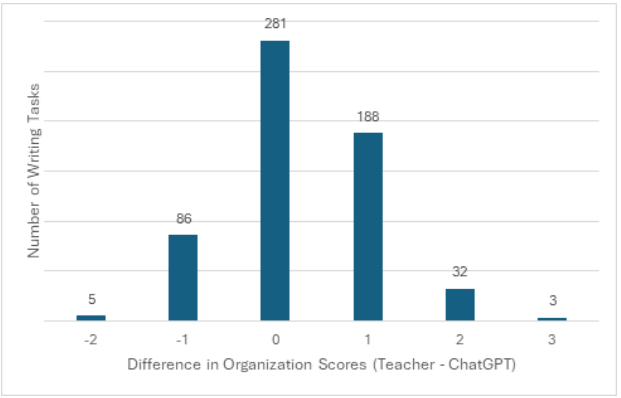

Figure 3. Difference in organisation scores (Teacher – ChatGPT)

Figure 4 shows the scoring alignment between ChatGPT and the teacher for the language part. It is seen that while 319 tasks were identical, the teacher assigned higher scores more frequently, particularly with 1-point (206 tasks) and 2-point (25 tasks) differences. ChatGPT scored 1 point higher in 42 tasks and 2 points higher in 1 task, but these occurrences were less common.

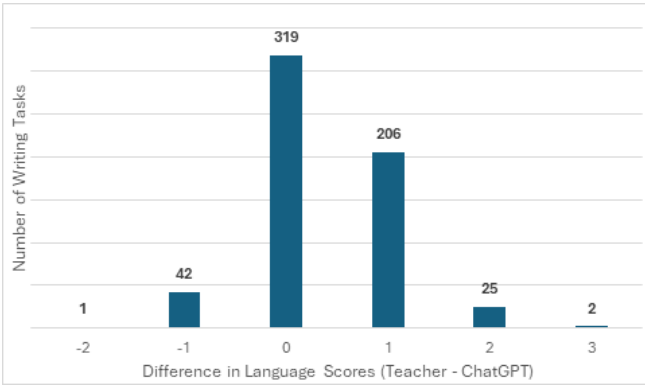

Figure 4. Difference in language scores (Teacher – ChatGPT)

To summarise, even though the teacher tended to give higher scores than ChatGPT, there was substantial agreement between them. For both pre- and post-feedback writing scores, there is a strong correlation between teacher and ChatGPT scores. This means that both scorers tend to produce similar results, with a high degree of alignment, and the observed correlation is not due to random chance, reflecting a genuine relationship.

At the end of the experimental data collection period, qualitative research was needed to support the research from the perspectives of the students. To achieve this, the teacher conducted an individual semi-structured interview, and 45 students participated in the interviews out of 53 students. In the interview, there were two major questions:

1. Which feedback do you prefer to correct your mistakes and improve your writing?
2. Why? Can you explain your reason by giving examples?

After analysing the interviews, for the first question, it was found that 24 students chose teacher feedback while 21 students chose ChatGPT feedback. Figure 5 illustrates the percentage of students' preferences for feedback type.

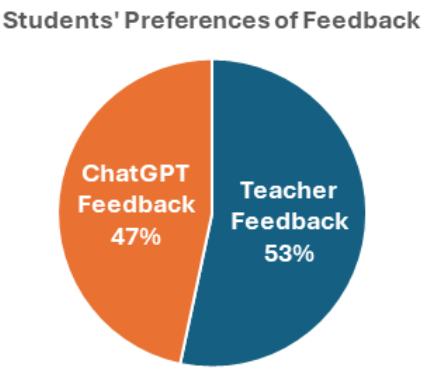

Figure 5. Students' preferences for feedback

For the second question, the students reflected on their reasons for choosing a feedback type. The comments of the students were recorded through note-taking, and an inductive thematic analysis was carried out. The recurring words were coded, and the codes were categorised under four categories for both ChatGPT and teacher feedback.

Table 15 presents the students' reasons to prefer teacher feedback in four categories, how many times the reasons were mentioned in the interviews, and sample student comments mentioning those reasons. Students who chose teacher feedback mostly thought that it was more concise and understandable when compared to ChatGPT feedback. They also stated that it was easy to see their mistakes on the paper because there were fewer explanations in the teacher's feedback. They also preferred teacher feedback because they found the English level simpler than ChatGPT feedback. For example, the teacher wrote sentences like, "if you like horror movies, it can't be boring." While ChatGPT wrote, "👎 Check your reasons. If you think a movie type is boring or terrible, usually it means you don't like it." The student who gave the example said: "I don't know some words in the second feedback form, and I need to look up a dictionary to understand the feedback."

Table 15. Students' reasons to prefer teacher feedback

Code	Number of students	Student comments
Simplicity of English	12	"I can understand it since the English level is simple." "It's English is easier." "It's English matches my level perfectly."
Ease of the seeing mistakes	8	"I can easily see where I make mistakes." "It is easy to find my mistakes." "I can notice my mistakes easily."
Conciseness	8	"… and there are short and clear sentences." "…there is less explanation, it resembles a concise summary." "The feedback is concise and clear."
Ease of understanding	6	"It is easier to understand and…" "It is more understandable." "I can understand easily."

Table 16 presents the students' reasons for preferring ChatGPT feedback, categorised into four themes, along with the frequency of each reason mentioned during interviews and sample student comments. Students who favoured ChatGPT feedback stated that it was more explanatory. They noted that ChatGPT not only highlighted their mistakes but also provided examples and explanations to help correct them. Additionally, they appreciated that ChatGPT feedback included both the incorrect and corrected versions of their sentences, whereas teacher feedback typically addressed errors at the word level. For example, one student wrote, "It lives at home." The teacher corrected this to "at home," while ChatGPT rewrote the full sentence as, "It lives at home."

Students also valued that ChatGPT acknowledged their strengths in addition to their weaknesses, which they felt was less common in teacher feedback. For instance, ChatGPT offered praise such as, "You write about where the cat lives. Good job!"—a comment that is not present in the teacher's feedback. Finally, students found the use of emojis in ChatGPT feedback motivating and engaging, with examples like, "Keep practicing, you are doing well! 😊 🌟" and "Your ideas are well-organised👍 ".

In conclusion, slightly more students preferred teacher feedback, citing its simplicity and clarity, while nearly half of the students favoured ChatGPT feedback for its detailed explanations and motivational tone. The distribution of preferences, with 53% choosing teacher feedback and 47% choosing ChatGPT feedback, reflects a relatively balanced division among students. This suggests that neither method clearly outperformed the other in terms of student experience.

Table 16. Students' reasons to prefer ChatGPT feedback

Code	Number of students	Student comments
Better explanation	9	"It is more explanatory…"
		"…and it has more explanations."
		"It explains why I made mistakes better."
Use of emoji	8	"…the feedback also included emojis, it is more fun."
		"It uses emojis like smiley faces and stars, and it makes me more motivated."
		"It explains positive and negative things using emojis."
Showing both the correct and incorrect versions of the sentences	6	"…it shows both incorrect and correct sentences."
		"It shows how my sentences were supposed to be and the incorrect version at the same time."
		"It writes both the correct and incorrect version of the sentences side by side."
Stressing both the strengths and weaknesses of the writing	6	"…shows also the strengths and not just the weaknesses."
		"It doesn't just stress the mistakes, but also the positive sides."
		"I can also see the positive sides of my writing."

DISCUSSION AND CONCLUSION

This study investigated the effectiveness and perceived value of ChatGPT-4-generated feedback compared to traditional teacher feedback in the context of secondary EFL learners' writing development. Interpreted through a feedback-oriented theoretical lens, the findings suggest that both teacher feedback and ChatGPT-generated feedback can support revision, but they do so in different ways. From the perspective of Hattie and Timperley's (2007) model, both feedback types appeared capable of helping students reduce the gap between current and desired performance, as reflected in the improvement observed across several writing tasks. However, teacher feedback appeared to offer stronger pedagogical guidance overall, likely because it was better able to prioritise significant issues and respond to the communicative and developmental needs of the learner. From a feedback literacy perspective, the qualitative findings are especially important: students did not evaluate feedback only in terms of correctness but also in terms of understandability, emotional tone, and perceived personal relevance. This helps explain why teacher feedback was often preferred even when AI feedback was seen as fast, clear, and useful. Viewed through Tri-System Theory, ChatGPT feedback may function as an external cognitive resource that

supports revision, but it may also encourage cognitive surrender when learners accept AI-generated suggestions without sufficient evaluation (Shaw & Nave, 2026). They also argue that AI can function as an external cognitive system that may support reasoning but also promote cognitive surrender.

Both feedback sources were linked to significant gains in Weeks 3 and 4, indicating that some writing tasks or instructional conditions may have been especially supportive of improvement regardless of whether feedback came from the teacher or ChatGPT. At the same time, the differing weekly patterns deserve attention. Teacher feedback led to significant gains in Weeks 1 and 5, whereas ChatGPT feedback produced gains in Weeks 2 and 6. This suggests that the two feedback sources may facilitate writing development through partly different processes. Teacher feedback may have been more effective when students required concise, familiar, and proficiency-appropriate guidance, while ChatGPT may have been particularly useful when students benefited from more elaborated explanations or sentence-level reformulations. These patterns indicate that teacher and AI feedback should not be treated as interchangeable, even when both contribute positively to learning. Overall, teacher feedback appeared slightly more effective, which is consistent with earlier research emphasising the pedagogical depth and contextual sensitivity of human feedback (Steiss et al., 2024). In this respect, teacher feedback seemed better able to identify and prioritise the most important issues and provide clearer instructional direction, which remain areas where AI feedback is still limited. This point is especially important given Shaw and Nave's (2026) argument that users may become more confident in AI-supported judgements even when the AI output is flawed. In the context of writing feedback, this suggests that the apparent clarity and authority of AI-generated comments may encourage acceptance without sufficient evaluation, reinforcing the continuing importance of teacher mediation.

A moderate to strong correlation between teacher-assigned and ChatGPT-generated scores suggests promising alignment, supporting findings from Shi and Aryadoust's (2024) review on the reliability of AI-generated feedback for lower-level writing features. However, student perceptions revealed a consistent preference for teacher feedback. While AI responses were valued for their speed and clarity, teacher feedback was perceived as more personal, motivational, and context-aware, echoing findings by Escalante et al. (2023) who emphasised the relational trust and engagement fostered by human feedback.

These results contribute to the growing discourse on integrating generative AI into writing pedagogy. The evidence supports a hybrid feedback model, where AI complements teacher input. AI tools like ChatGPT are effective at providing descriptive and stylistic feedback (Banihashem et al., 2025), while teachers excel at identifying conceptual and rhetorical weaknesses. Guo et al. (2024) demonstrated that AI-supported peer review can enhance students' feedback literacy, positioning AI not just as a corrective tool but as a scaffold for developing evaluative skills. Strategically implemented, AI can reduce teacher workload and increase the frequency of formative feedback, especially valuable in large or under-resourced classrooms. However, the pedagogical limitations of AI-generated feedback

must be critically considered. Recent work suggests that AI may encourage uncritical reliance under some conditions (Shaw & Nave, 2026). AI systems cannot often interpret nuanced rhetorical choices, cultural context, and learner intent (Alawad & Hamid, 2025). Moreover, overreliance on AI may risk reducing feedback to a transactional exchange, undermining the dialogic and developmental nature of writing instruction (Ting et al., 2025). The discursive stance adopted by AI, whether prescriptive or facilitative, can shape learner agency and perceptions of authority in unintended ways.

Emotional and cognitive engagement with feedback is another critical dimension. Studies show that while AI-generated feedback can enhance motivation and reduce anxiety, it may lack the emotional resonance necessary to foster deep learning. Zhang and Wang (2025) found that hybrid feedback models improved cognitive and affective engagement, but behavioural engagement remained superficial without teacher scaffolding. Similarly, Zhan and Yan (2025) emphasised the need to cultivate feedback literacy including emotional reflexivity and evaluative judgement to ensure meaningful uptake of AI-generated feedback.

These findings suggest several directions for future research. Longitudinal studies could assess the sustained impact of AI-generated feedback on writing development. Qualitative investigations into students' revision behaviours and feedback uptake would offer deeper insight into learning processes. Cross-linguistic and cross-cultural studies are also needed, particularly in multilingual EFL contexts where AI training data may encode cultural or linguistic biases (Alawad & Hamid, 2025). Additionally, professional development for teachers should include both technical training and pedagogical frameworks for integrating AI meaningfully into instruction.

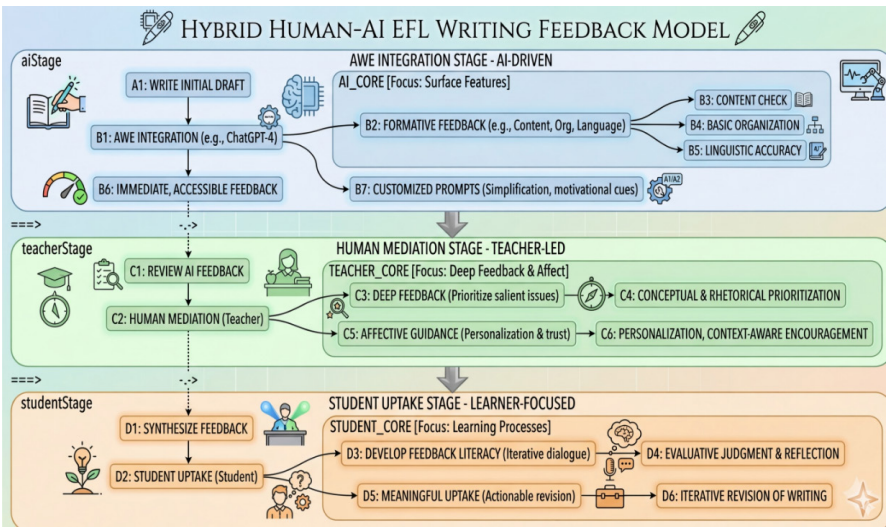

Figure 2. Hybrid human-AI EFL writing feedback model(Note: The figure was created by using Gemini).

Figure 2 illustrates the Hybrid Human-AI EFL Writing Feedback Model. It has a sequential, three-stage feedback loop for secondary EFL writing. The process begins with an AI-driven stage, where Automated Writing Evaluation (AWE) tools like ChatGPT provide immediate formative feedback focusing primarily on surface-level features such as content, basic organisation, and linguistic accuracy, tailored with personalised prompts for the A1–A2 proficiency level. The next stage is Human Mediation, where the teacher reviews the AI output and provides deeply personalised, context-aware, and emotionally engaging feedback, prioritising conceptual and rhetorical nuances that AI often overlooks to prevent student "cognitive surrender." The final stage is Student Uptake, which is learner-focused and centered on developing students' feedback literacy. Through iterative dialogue and reflective processes, students critically synthesise and act upon the integrated feedback, leading to evaluative judgement, meaningful engagement, and multiple iterative revisions of their final writing drafts.

The study contributes to broader research on feedback and AI-supported language learning in three ways. First, it extends the emerging literature on AI-generated feedback beyond higher education by focusing on secondary school EFL learners, a population that remains underrepresented in current research. Second, the findings show that correlation between teacher and AI scores should not be interpreted as evidence of pedagogical equivalence. Although ChatGPT-generated feedback showed substantial alignment with teacher scoring, students still distinguished clearly between the two forms of feedback in terms of clarity, personalisation, motivational value, and ease of use. Third, the study reinforces a broader shift in feedback research from viewing feedback as information transmission alone toward understanding it as a process of learner engagement. In AI-supported language learning, this means that the value of AI feedback depends not only on technical quality but also on whether learners can meaningfully interpret and act on it. For this reason, the findings support a hybrid model in which AI strengthens access to timely formative feedback, while teachers remain central to contextualisation, emotional support, and deeper pedagogical judgement. In sum, while GenAI showed considerable potential as a feedback tool, it did not surpass teacher feedback in improving student writing or earning student trust. Its value lies in its role as a complementary resource, enhancing efficiency and reach when used judiciously. A carefully calibrated human-AI partnership in feedback provision may support the development of more autonomous, reflective writers. Realising this potential, however, requires ongoing empirical validation, critical reflection, and a commitment to pedagogical integrity in the age of intelligent automation.

Practical Suggestions for EFL Practitioners Using ChatGPT Feedback

To support effective integration of AI-generated feedback in EFL writing instruction, this study offers several practical recommendations grounded in classroom experience and empirical findings. First, AI-generated feedback is best utilised during early drafting stages, where it can efficiently address surface-level issues such as grammar, sentence structure, and clarity. Teachers, in turn, can focus their efforts on final revisions, providing higher-order feedback related to argumentation, coherence, and creativity.

It is essential to cultivate students' AI-generated feedback literacy. Learners should be guided to critically evaluate GenAI's suggestions rather than accept them uncritically. The concept of cognitive surrender helps explain why learners may accept AI-generated feedback too quickly (Shaw & Nave, 2026). Encouraging comparison with teacher feedback and reflection on usefulness fosters metacognitive awareness and strengthens digital literacy. To ensure accessibility for lower-proficiency learners, prompts used with GenAI should be adapted to their linguistic level. For example, simplified language, short sentences, and visual cues can make AI-generated feedback more engaging and comprehensible for A1–A2 students.

While GenAI offers efficiency, it cannot address the emotional and affective dimensions of learning. Therefore, AI-generated feedback should be complemented with pedagogical scaffolding, such as teacher check-ins, peer discussions, or reflective activities that support motivation and emotional engagement. Moreover, although AI-generated scores may align moderately with teacher assessments, they can misprioritise feedback or overlook deeper content issues. For this reason, GenAI should be used as a formative support tool rather than a summative assessment method.

To discourage overreliance on AI and promote authentic writing, educators should design tasks that require personal experience, classroom-based input, or visual analysis contexts that are difficult for AI to replicate meaningfully. Establishing clear ethical guidelines for AI use is also critical. Classroom policies should address academic integrity, authorship, and responsible technology use, helping students develop ethical awareness from an early age.

The dual-feedback design employed in this study can be adapted as a classroom activity. Students can compare AI and teacher feedback, evaluate their respective strengths and limitations, and make informed decisions about which suggestions to implement. Involving students in the design of GenAI prompts further promotes ownership and deepens their understanding of constructive feedback. Finally, in contexts with limited digital access, teachers can mediate AI-generated feedback by transcribing and distributing it in print. While this requires careful planning, it remains a viable strategy for ensuring equitable access to individualised support.

ACKNOWLEDGEMENT

The second author contributed as the Master's dissertation supervisor, offering scholarly guidance, supervision, and support during the development of the study and manuscript.

REFERENCES

Alawad, E. A., & Hamid, F. A. (2025). An interdisciplinary review and synthesis of applied linguistics in translation studies: Bridging gaps and advancing research. *Architecture Image Studies, 6*(3), 1246–1257. https://doi.org/10.62754/ais.v6i3.436

Alzahrani, N. (2026). AI-First Critique Learning (AFCL): A framework for restoring assessment integrity in the age of Generative AI. *International Journal of AI in Pedagogy, Innovation, and Learning Futures, 2026*(1), 1–23. https://doi.org/10.46787/ijaipil.v2026i1.6945

Agustini, N. P. O. (2023). Examining the role of ChatGPT as a learning tool in promoting students' English language learning autonomy relevant to Kurikulum Merdeka Belajar. *Edukasia: Jurnal Pendidikan Dan Pembelajaran, 4*(2), 921–934. https://doi.org/10.62775/edukasia.v4i2.373

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing, 57*, 100745. https://doi.org/10.1016/j.asw.2023.100745

Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education, 21*(1), 23. https://doi.org/10.1186/s41239-024-00455-4

Banihashem, S. K., Bond, M., Bergdahl, N., Khosravi, H., & Noroozi, O. (2025). A systematic mapping review at the intersection of artificial intelligence and self-regulated learning. *International Journal of Educational Technology in Higher Education, 22*(1), 50. https://doi.org/10.1186/s41239-025-00548-8

Baskara, F. R. (2023). Integrating ChatGPT into EFL writing instruction: Benefits and challenges. *International Journal of Education and Learning, 5*(1), 44–55. https://doi.org/10.31763/ijele.v5i1.858

Bewersdorff, A., Zhai, X., Roberts, J., & Nerdel, C. (2023). Myths, mis- and pre-conceptions of artificial intelligence: A review of the literature. *Computers and Education: Artificial Intelligence, 4*, 100143. https://doi.org/10.1016/j.caeai.2023.100143

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. https://doi.org/10.4324/9780203832400

Casal, J. E., & Kessler, M. (2023). Can linguists distinguish between ChatGPT/AI and human writing? A study of research ethics and academic publishing. *Research Methods in Applied Linguistics, 2*(3), 100068. https://doi.org/10.1016/j.rmal.2023.100068

Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education, 20*, 38. https://doi.org/10.1186/s41239-023-00408-3

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International, 61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). SAGE.

Crosthwaite, P., Ningrum, S., & Lee, I. (2022). Research trends in L2 written corrective feedback: A bibliometric analysis of three decades of Scopus-indexed research on L2 WCF. *Journal of Second Language Writing, 58*, 100934. https://doi.org/10.1016/j.jslw.2022.100934

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y.-S., Gašević, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. In *Proceedings of the 2023 IEEE International Conference on Advanced Learning Technologies (ICALT)* (pp. 323–325). IEEE. https://doi.org/10.1109/ICALT58122.2023.00100

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*, 57. https://doi.org/10.1186/s41239-023-00425-2

Evmenova, A. S., Regan, K., Mergen, R., & Hrisseh, R. (2024). Improving writing feedback for struggling writers: Generative AI to the rescue? *TechTrends, 68*(4), 790–802. https://doi.org/10.1007/s11528-024-00965-y

Farrokhnia, M., Latifi, S., Papadopoulos, P. M., Hogenkamp, L., Gijlers, H., Khosravi, H., & Noroozi, O. (2026). Generative AI offers more, but students revise less: Comparing the effects of teacher and AI feedback on student essay revisions. *International Journal of Educational Technology in Higher Education, 23*(1), 6. https://doi.org/10.1186/s41239-026-00579-9

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies, 15*(1), 6. https://doi.org/10.3390/soc15010006

Gunes, A., & Liman Kaban, A. (2025). A Delphi study on ethical challenges and ensuring academic integrity regarding AI research in higher education. *Higher Education Quarterly, 79*(4), e70057. https://doi.org/10.1111/hequ.70057

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies, 29*(7), 8047–8073. https://doi.org/10.1007/s10639-023-12146-0

Guo, K., Pan, M., Li, Y., & Lai, C. (2024). Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability. *The Internet and Higher Education, 63*, 100962. https://doi.org/10.1016/j.iheduc.2024.100962

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

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14–26. https://doi.org/10.3102/0013189X033007014

Kang, E., & Han, Z. (2015). The efficacy of written corrective feedback in improving L2 written accuracy: A meta-analysis. *The Modern Language Journal, 99*(1), 1–18. https://doi.org/10.1111/modl.12189

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(2), 537–550. https://doi.org/10.1177/00336882231162868

Kostka, I., & Toncelli, R. (2023). Exploring applications of ChatGPT to English language teaching: Opportunities, challenges, and recommendations. *TESL-EJ, 27*(3), n3. https://doi.org/10.55593/ej.27107int

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X. -H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv.* https://arxiv.org/abs/2506.08872

Lam, C., & Wolfe, J. (2023). An introduction to quasi-experimental research for technical and professional communication instructors. *Journal of Business and Technical Communication, 37*(2), 174–193. https://doi.org/10.1177/10506519221143111

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

Moorhouse, B. L., Wong, K. M., & Li, L. (2023). Teaching with technology in the post-pandemic digital age: Technological normalisation and AI-induced disruptions. *RELC Journal, 54*(2), 311–320. https://doi.org/10.1177/00336882231176929

OpenAI. (2023). GPT-4 technical report. https://arxiv.org/abs/2303.08774

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

Rassaei, E., & Jabbarpoor, S. (2025). Effects of textual enhancement on L2 development: A meta-analysis. *International Review of Applied Linguistics in Language Teaching.* https://doi.org/10.1515/iral-2025-0118

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching, 6*(1), 342–363. https://doi.org/10.37074/jalt.2023.6.1.9

Sharples, M. (2025). A systems approach to AI and education in a post-digital world. *Theory Into Practice, 64*(4), 483–491. https://doi.org/10.1080/00405841.2025.2528549

Shaw, S. D., & Nave, G. (2026). Thinking-fast, slow, and artificial: How AI is reshaping human reasoning and the rise of cognitive surrender. *SSRN*. https://doi.org/10.2139/ssrn.6097646

Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL, 36*(2), 187–209. https://doi.org/10.1017/S0958344023000265

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Stewart, A. E., Rao, A., Michaels, A., Sun, C., Duran, N. D., Shute, V. J., & D'Mello, S. K. (2023, June). CPSCoach: The design and implementation of intelligent collaborative problem solving feedback. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova. (Eds.), *International conference on artificial intelligence in education* (pp. 695–700). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36272-9_58

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing, 57*, 100752. https://doi.org/10.1016/j.asw.2023.100752

Su, J., Guo, K., Chen, X., & Chu, S. K. W. (2024). Teaching artificial intelligence in K-12 classrooms: A scoping review. *Interactive Learning Environments, 32*(9), 5207–5226. https://doi.org/10.1080/10494820.2023.2212706

Teng, M. F. (2024). "ChatGPT is the companion, not enemies": EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. *Computers and Education: Artificial Intelligence, 7*, 100270. https://doi.org/10.1016/j.caeai.2024.100270

Ting, L., Singh, C. K. S., Kiong, T. T., Rakhimjonov, F., & Singh, T. S. M. (2025). Mentor or examiner? A critical discourse analysis of AI-generated feedback in EFL writing education. *International Journal of Academic Research in Progressive Education & Development, 14*(4), 664–684. https://doi.org/10.6007/IJARPED/v14-i4/26651

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

von Garrel, J., & Mayer, J. (2023). Artificial intelligence in studies-use of ChatGPT and AI-based tools among students in Germany. *Humanities and Social Sciences Communications, 10*, Article 1. https://doi.org/10.1057/s41599-023-02304-7

Wang, W. S., Lin, C. J., Lee, H. Y., Huang, Y. M., & Wu, T. T. (2025). Integrating feedback mechanisms and ChatGPT for VR-based experiential learning: Impacts on reflective thinking and AIoT physical hands-on tasks. *Interactive Learning Environments, 33*(2), 1770–1787.

Xiao, F., Zhu, S., & Xin, W. (2026). Exploring the landscape of generative AI (ChatGPT)-powered writing instruction in English as a foreign language education: A scoping review. *ECNU Review of Education, 9*(1), 1–19. https://doi.org/10.1177/20965311241310881

Yamashita, T. (2025). Exploring potential biases in GPT-4o's ratings of English language learners' essays. *Language Testing, 42*(3), 344–358. https://doi.org/10.1177/02655322251329435

Yoon, S. Y., Miszoglad, E., & Pierce, L. R. (2023). Evaluation of ChatGPT feedback on ELL writers' coherence and cohesion. arXiv:2310.06505. *arXiv Preprint.*

Zhan, Y., & Yan, Z. (2025). Students' engagement with ChatGPT feedback: Implications for student feedback literacy in the context of generative artificial intelligence. *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2025.2471821

Zhang, J., & Wang, J. (2025). Student perceptions of hybrid feedback: Using Gen-AI to enhance engagement with EAP writing feedback. *The JALT CALL Journal, 21*(2), 2175–2175. https://doi.org/10.29140/jaltcall.v21n2.2175

Zheng, Y., Yu, S., & Liu, Z. (2023). Understanding individual differences in lower-proficiency students' engagement with teacher-written corrective feedback. *Teaching in Higher Education, 28*(2), 301–321. https://doi.org/10.1080/13562517.2020.1806225

Zou, S., Guo, K., Wang, J., & Liu, Y. (2025). Investigating students' uptake of teacher- and ChatGPT-generated feedback in EFL writing: A comparison study. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2024.2447279