
Human–AI collaboration in academic writing instruction: A calibrated feedback model for EAP

Reza Farzi
University of Ottawa

Abstract

This study adopts a mixed-methods design to compare feedback on student essays generated by a human instructor and ChatGPT-3.5. It examines scoring consistency, students' perceptions of feedback usefulness, and alignment with instructional objectives. The study draws on two argumentative writing tasks completed by 30 students in a Canadian English for Academic Purposes (EAP) course and assessed independently by the instructor and ChatGPT-3.5. The findings indicate that ChatGPT-3.5 identifies surface-level issues and produces scores closely aligned with the instructor's but is less effective at addressing rhetorical structure and contextual nuance. Students expressed a clear preference for a hybrid feedback model that combines the immediacy of AI-generated comments with the pedagogical value of instructor feedback. The findings also suggest that through calibrated prompts and pedagogical (re-)framing, generative AI can effectively supplement, but not replace, human feedback. The study offers guidance for integrating generative AI into academic writing instruction while preserving instructional quality and student agency.

Key words: generative AI; automated written feedback; ChatGPT; English for Academic Purposes (EAP); student perceptions

Résumé

Cette étude adopte une méthodologie mixte pour comparer la rétroaction d'un enseignant et de ChatGPT-3.5 sur des écrits étudiants. Elle examine la cohérence de la notation, l'utilité perçue de la rétroaction et son alignement sur les objectifs pédagogiques. Les données proviennent de deux séries de textes argumentatifs rédigés par 30 étudiants d'un cours canadien d'anglais sur objectifs universitaires (EAP), évalués par l'enseignant et ChatGPT-3.5. Les résultats montrent que ChatGPT-3.5 repère les erreurs de surface et produit des notes proches de celles de l'enseignant, mais traite moins efficacement la structure rhétorique et les nuances contextuelles. Les étudiants préférèrent un modèle hybride combinant l'immédiateté de la rétroaction générée par l'IA et la valeur

pédagogique de celle de l'enseignant. Avec des requêtes calibrées et un recadrage pédagogique, l'IA générative peut compléter, sans remplacer, la rétroaction humaine. L'étude propose des pistes pour intégrer l'IA générative à l'enseignement de l'écriture universitaire tout en préservant qualité pédagogique et agentivité étudiante.

Mots-clés : IA générative ; rétroaction écrite automatisée ; ChatGPT ; anglais sur objectifs universitaires (EAP) ; perceptions des étudiants

Introduction

The integration of generative artificial intelligence (GenAI) into education has prompted global debate about the future of teaching, learning, and assessment (Ogunleye et al., 2024; Samala et al., 2024). In particular, the widespread adoption of large language model (LLM) systems such as OpenAI's ChatGPT has begun to reshape how writing instruction and feedback are mediated (Gayed et al., 2022; Liang et al., 2025). Recent international policy discussions, including the *OECD Digital Education Outlook 2026* (OECD, 2026), have further emphasized the need for ethically grounded and pedagogically meaningful uses of GenAI in educational settings. Trained on extensive linguistic data, these systems can generate contextually appropriate text across genres and offer immediate evaluative and revision-oriented responses, leading educators to reconsider established feedback practices (Harunasari, 2023). In second language (L2) writing instruction, timely and effective feedback plays a crucial role in learner development (Ferris, 2014). In English for Academic Purposes (EAP) programs, students must master complex rhetorical patterns and discipline-specific conventions. In this context, feedback not only addresses errors but also fosters higher-order thinking and greater awareness of academic discourse (Hyland, 2016; Hyland & Hyland, 2019). Human instructors can bring pedagogical insight, empathy, and contextual knowledge to the feedback process, but time pressures and institutional workloads often limit their capacity to do so. GenAI tools can assist instructors by providing timely and scalable responses; however, questions remain about their effectiveness and appropriateness for teaching purposes (Shi & Aryadoust, 2024).

A central challenge in using GenAI tools such as ChatGPT is aligning AI-generated feedback with instructional goals. Without proper calibration, AI-generated responses tend to focus on sentence-level features such as grammar and vocabulary and devote less attention to coherence, idea development, and genre expectations (Bender & Koller, 2020; Farzi, 2024). Students may also rely too heavily on GenAI-generated suggestions or have difficulty interpreting abstract or unclear feedback (Nelson et al., 2025). These challenges underscore the importance of structured guidance for GenAI to function as a useful instructional resource.

Building on these concerns, this study examines whether calibration strategies, such as prompt customization and the use of exemplars, can better align GenAI outputs with instructional goals in an EAP setting (Farzi, 2024). Drawing on principles from second language pedagogy and established assessment practices, the study proposes an empirically grounded approach to incorporating calibrated GenAI feedback

into academic writing instruction. Specifically, it addresses the following research questions:

1. To what extent can calibrated GenAI feedback align with human instructor feedback in domain-specific scoring and overall writing assessment?
2. How do students perceive the usefulness, clarity, and instructional quality of AI feedback relative to traditional instructor feedback?
3. What are the pedagogical implications of integrating AI and instructor feedback within a collaborative instructional model?

In response to these research questions, the study adopts a mixed-methods design that combines quantitative comparisons of scoring outcomes with qualitative analysis of written feedback. It examines assessment data alongside student perspectives to develop practice-oriented insights into the use of GenAI in L2 writing contexts. The findings are intended to inform decision-making in EAP programs and clarify the role of GenAI as a complement to, rather than a substitute for, instructor-led pedagogy.

Literature review

Automated writing evaluation (AWE) tools have a long history in L2 pedagogy and language assessment. Early systems such as Project Essay Grade (Page, 2003) and Intelligent Essay Assessor (Landauer, 2003) relied on statistical and computational models to evaluate features of student writing (Dikli, 2006). These tools had limited capacity to evaluate discourse-level elements such as argument coherence and genre conventions. With advances in natural language processing and neural networks, more sophisticated tools such as ETS's Criterion incorporated domain-specific scoring and holistic evaluation; however, their feedback remained largely formulaic and lacked contextual sensitivity (Cotos, 2011).

The recent emergence of LLMs such as ChatGPT, Copilot, and Gemini, all built on transformer architecture (Vaswani et al., 2017), represents a significant expansion of GenAI capabilities. These models generate human-like text responses by modelling complex semantic relationships across large textual corpora (Devlin et al., 2018). Educators and researchers have taken a keen interest in these tools because of their responsiveness, prompt adaptability, and linguistic fluency. Despite their potential to provide real-time, grammatically accurate feedback, concerns remain about their reliability, consistency, and educational value (Fleckenstein et al., 2023; Radford et al., 2019; Rana et al., 2026).

Research on the use of GenAI feedback in writing instruction has produced mixed results. Some studies highlight benefits, including improved revision habits, greater learner independence, support for multilingual students in identifying basic errors, reduced writing anxiety, and improved fluency and accuracy (Farzi, 2024; Gayed et al., 2022; Wang, 2024; Zhang et al., 2025). In classroom settings, Lee and Moore (2024) found that students were more engaged when GenAI feedback supported, rather than replaced, instruction. Similarly, a systematic review by Shi and Aryadoust (2024) reported generally positive learner outcomes associated with AWE tools. These findings suggest that, when used appropriately, GenAI can support students during early drafting and self-editing.

Conversely, L2 education researchers have raised several critical concerns. Zhai et al. (2024) highlight learners' tendency to rely excessively on GenAI suggestions, often resulting in mechanical and less reflective revision. Similarly, Nelson et al. (2025) reported that students sometimes misunderstand general GenAI feedback or do not apply suggestions when comments are insufficiently contextualized. These findings align with earlier critiques that, although AI-generated feedback is fast and useful, it focuses primarily on surface-level correctness and devotes less attention to higher-order rhetorical concerns. Concerns about fairness also remain. For example, students from different linguistic backgrounds or underrepresented groups may receive inconsistent feedback that does not align with their specific needs (Shi & Aryadoust, 2024). LLMs are trained on corpora shaped by dominant cultural and discourse conventions, which may not adequately represent these learners' linguistic identities (Bender & Koller, 2020; Evmenova et al., 2024; Farzi, 2024). Consequently, their language use may be misinterpreted or evaluated against inappropriate norms.

L2 writing experts have long emphasized feedback that supports both cognitive and metacognitive growth (Ferris, 2014; Hyland, 2016; Hyland & Hyland, 2019). Human instructors not only correct errors but also help students make rhetorical choices, consider their audience, and develop an academic voice. When used alone, GenAI feedback lacks this pedagogical scaffolding. Accordingly, some researchers have proposed hybrid models in which GenAI feedback is combined with instructor guidance to ensure that feedback is both timely and pedagogically meaningful (Farzi, 2024; Shi & Aryadoust, 2024; Zhang et al., 2025). These models support an ecological view of feedback that considers how the tool, learner, and instructor interact throughout writing development. Recent research suggests that students benefit most from GenAI when they engage with it interactively and revise iteratively rather than use it simply as a source of information. This finding points to the value of guided, reflective AI use in higher education writing contexts (A. Nguyen et al., 2024).

Recent empirical research has identified several strategies for improving the quality and alignment of AI-generated feedback. In a controlled EAP study, Farzi (2024) showed that using domain-specific rubrics and model answers for prompt calibration improved consistency between ChatGPT and instructor scoring. Through iterative prompt refinement and the use of anchor texts, the model aligned more closely with instructor evaluations of abstract aspects of writing, particularly coherence and development. Importantly, calibration increased interrater reliability between the instructor's scores and ChatGPT's scores, thereby reducing differences between human and AI ratings. These findings indicate that, when combined with appropriate pedagogical approaches, AI tools may function as reliable support systems in assessment contexts.

Building on this evidence, Farzi (2024) introduced a model in which GenAI is embedded thoughtfully within instructional practice rather than used as an isolated tool. In this approach, GenAI was not treated as an independent evaluator; instead, it formed part of a structured feedback process that began with instructor-designed rubrics and calibration. Instructor involvement in prompt design and response align-

ment made ChatGPT's feedback more contextually appropriate and closely aligned with pedagogical goals. Rather than positioning AI and instructor feedback as alternatives, the model treated them as complementary. GenAI addressed mechanical and language-level issues, whereas the instructor focused on content and higher-order meaning. This collaboration helped maintain students' trust in the feedback and preserved the pedagogical relationship that supports writing development. At the same time, recent meta-analytic research suggests that AI-generated feedback can sometimes produce learning outcomes comparable to those associated with human feedback (Kaliisa et al., 2026).

Despite these developments, ethical and practical challenges remain in the use of AI in L2 writing instruction. These challenges include privacy concerns, opacity in AI decision-making, and the risk of homogenizing student writing when AI is used without critical supervision (Bender & Koller, 2020; García-López & Trujillo-Liñán, 2025). In addition, the boundary between assisting students and completing work for them remains unclear, raising questions about academic integrity (Lund et al., 2025). Instructors must manage these issues through clear policies, reflective practice, and regular monitoring of student progress. These concerns require a comprehensive approach that respects ethical principles and helps students use the technology responsibly and critically (van Niekerk et al., 2025).

As research on GenAI in education grows, practical studies in authentic L2 classrooms remain limited. Many existing studies discuss GenAI broadly or focus primarily on its technical features. Evidence is particularly limited on how GenAI functions in classroom settings, including its relationship to classroom dynamics, feedback practices, and trust-building with learners.

The instructor's role in this context also remains understudied. Although growing attention is being paid to how GenAI can support writing and revision, less is known about how instructors use these tools in their teaching. Few studies provide concrete examples of how instructors connect GenAI tools with instructional goals and student needs. The present study addresses this gap by examining how the instructor calibrates GenAI feedback and how students perceive its value for their learning.

Methodology

This study was conducted in a first-year EAP writing course at a Canadian university. The course was designed to prepare undergraduate students for writing tasks across academic disciplines, and most participants used English as an L2 or as an additional language. A total of 30 students¹ from a range of linguistic and cultural backgrounds participated. They were placed in the course on the basis of one of three measures: results from the university's intensive language program, institutional placement testing, or standardized external examinations. For most participants, the course fulfilled a university language requirement. Most participants' proficiency ranged from mid- to upper-intermediate, approximately equivalent to IELTS bands 6–6.5.

¹ Ethics approval for the study was granted by the University of Ottawa's Research Ethics Board (REB). All participants provided informed consent and were assured that participation would not affect their grades.

The study used a mixed-methods approach, as described by Creswell and Plano Clark (2017), combining quantitative score analysis with a close qualitative examination of written feedback. This design was selected because quantitative analysis alone could not capture students' interpretations and perceptions of feedback, whereas qualitative analysis alone could not establish scoring alignment or interrater reliability between human and AI raters.

The dataset comprised 60 student essays. Each student submitted two writing tasks during the term: an argumentative essay and an argumentative synthesis. These assignments were selected not only because they are common in academic settings but also because each requires distinct forms of reasoning and organization (Hyland, 2016; Hyland & Hyland, 2019).

The departmental rubric used in the study (see [Appendix A](#)) drew on established frameworks in L2 writing assessment (Uludag & McDonough, 2022). It comprised four categories: (1) lexical accuracy and range, (2) grammatical accuracy and range, (3) organization and development of the topic, and (4) overall effectiveness. Each category was evaluated on a 5-point scale, on which 1 indicated poor performance and 5 indicated performance that exceeded expectations. The descriptors predated the study and had been used and discussed in previous terms during professional development sessions in which instructors rated sample essays and revised the rubric language to improve agreement.

The lexical category (1) assessed academic and figurative vocabulary, word-form accuracy, spelling, collocations, and overall word choice. At higher score levels, the writing demonstrated a wide range of precise vocabulary, including discipline-specific and idiomatic expressions used appropriately in context. Grammatical accuracy (2) encompassed the variety and control of simple and complex structures, including subject-verb agreement, tense use, and advanced constructions such as conditionals and embedded clauses. The highest-rated writing demonstrated grammatical flexibility and accuracy with minimal errors. For organization and development (3), raters assessed the logical arrangement of ideas, use of cohesive devices, and extent to which explanations, reasons, and details supported the argument. Stronger responses demonstrated cohesion and coherence from introduction to conclusion and included relevant, specific supporting content. Overall effectiveness (4) assessed how clearly and naturally the writing communicated its message and how much effort was required from the reader to follow it. Higher-rated responses were clear, fluent, and easy to understand, with few instances of awkward phrasing or unclear transitions. In previous terms, the rubric categories had also been compared with national EAP benchmarks to assess alignment. During regular departmental workshops, instructors collaboratively reviewed anonymized student work to build a shared understanding of the criteria and improve scoring reliability.

The course instructor and ChatGPT-3.5 (OpenAI, 2023) independently assessed each essay. ChatGPT-3.5 was selected because it was the version available at the time of data collection and reflected a realistic classroom context. The instructor had substantial experience in EAP writing assessment. The ChatGPT evaluations occurred across three calibration phases (see [Appendix B](#)). In Phase 1, ChatGPT-3.5 received the original rubric and a basic prompt directing it to assign a score for

each category; no additional clarification was provided. In Phase 2, the prompt was revised to include definitions and elaborations for the rubric's four criteria. Each criterion was supported by simple examples intended to guide the model's interpretation. In Phase 3, annotated sample essays representing high, medium, and low performance across the rubric criteria were embedded in the prompt as reference points. This incremental calibration made it possible to examine how greater specificity affected alignment between ChatGPT's scores and human judgment.

The calibration process was informed by Farzi (2024), who showed that example responses combined with detailed instructional prompts improved ChatGPT's alignment with human scoring. The course instructor, who had expertise in writing assessment, created the prompts. A key feature of the design was the use of anchor texts, which helped the model interpret and apply the rubric across different genres. This method also reflects broader assessment literature, in which benchmark texts and rater training are considered important for maintaining construct validity and score reliability (Barkaoui, 2010).

The instructor, who had more than ten years of EAP teaching and assessment experience, independently rated all 60 essays without access to the AI-generated scores. This blinded scoring procedure was intended to minimize potential bias, consistent with recommended practices in educational measurement (Shi & Aryadoust, 2024).

After scoring, all 30 participants completed a 16-item survey comprising 12 closed-ended questions and four open-ended questions (see [Appendix C](#)). The survey addressed patterns of AI use, perceived effects on language learning, and students' views of AI-generated feedback and took approximately 15–20 minutes to complete. The open-ended responses were analyzed thematically using the six-step method described by Braun and Clarke (2006): (1) familiarization with the data, (2) generation of initial codes, (3) identification of themes, (4) review of themes, (5) definition and naming of themes, and (6) production of the report. The researcher manually coded the responses and conducted a second review to improve consistency. All student comments were anonymized, and the review continued until it yielded no new themes. Themes included perceived clarity and usefulness, trust in the feedback, and motivation to revise. The instructor also wrote a reflective memo describing their experience of comparing their feedback with ChatGPT's.

For the quantitative analysis, paired t-tests were used to examine the differences between the scores assigned by the human and AI raters. Cohen's *d* quantified the magnitude of those differences, and Cohen's kappa assessed agreement between the raters across calibration stages. These statistics are well-established in language assessment research for comparing scoring consistency and reliability (McNamara, 2000).

Results

The study compared AI-generated feedback with feedback provided by a human instructor in an EAP academic writing context. Both quantitative score data and qualitative written feedback were analyzed.

Quantitative findings

To compare AI-generated and human scores, a paired samples t-test was conducted on the total scores from both raters. The results are summarized in Table 1.

Table 1
Comparison of human and AI scores across essays

Group	Sample size (<i>n</i>)	Mean (<i>M</i>)	Standard deviation (<i>SD</i>)	<i>t</i> -value (<i>t</i>)	Degrees of freedom (<i>df</i>)	<i>p</i> -value (<i>p</i>)	Effect size (Cohen's <i>d</i>)
Human instructor	60	74.23	4.50	1.47	118	.104	0.27
ChatGPT-3.5	60	72.98	4.68				

Note. Scores were based on a four-domain analytic rubric. A paired *t*-test indicated no statistically significant difference ($p > .05$).

As shown in Table 1, the human instructor assigned slightly higher scores ($M = 74.23$) than ChatGPT-3.5 ($M = 72.98$). The *p*-value of .104 indicates that this difference was not statistically significant. The small effect size (Cohen's $d = 0.27$) suggests that the difference may have limited practical importance. This result is consistent with Farzi (2024), who reported close alignment between calibrated AI scores and human scores.

This similarity suggests that ChatGPT-3.5 can produce scores that are in line with human evaluations, making it a useful support tool for formative writing assessment. In formative assessment, scoring consistency helps ensure that feedback is fair, reliable, and based on clear criteria, which is essential for guiding student improvement. Since ChatGPT's scores aligned closely with those of the human evaluator in this study, it may be able to provide timely feedback that supports instructors in promoting students' writing development. These findings are broadly consistent with earlier research suggesting that AI can help standardize formative feedback without sacrificing fairness or reliability (Farzi, 2024; Nelson et al., 2025; Shi & Aryadoust, 2024).

Figure 1 shows the distributions of scores assigned by the human instructor and ChatGPT-3.5. The median scores are close, and the interquartile ranges show similar dispersion. Neither set includes noticeable outliers, suggesting that the raters applied the rubric consistently. The visual pattern is consistent with the statistical test, supporting the notion that ChatGPT's scoring is systematic and reasonably aligned with human evaluation. In this context, "systematic" refers to the model's tendency to apply scoring criteria predictably and consistently across comparable responses. Given these similarities, the AI model seems to offer a dependable option for supporting formative assessment in writing courses, particularly when instructors need quick feedback without compromising fairness or scoring consistency.

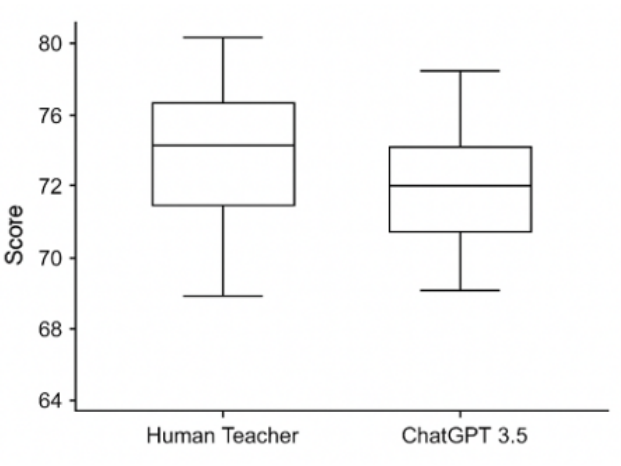


Figure 1: Distribution of Human and AI Scores

Note. Boxplot comparing total writing scores assigned by a human instructor and ChatGPT-3.5 across 60 essays. Both distributions appear approximately symmetric, with similar interquartile ranges and medians.

Qualitative findings

In addition to the score data, open-ended student survey responses and the instructor’s reflective memo were analyzed thematically. Table 2 summarizes the frequency of the key themes identified in the data.

Table 2
Frequency of student feedback themes (N = 30)

Theme	Description	Frequency
Clarity and specificity	Mentions of the clarity of grammar- or vocabulary-focused feedback	25
Timeliness and accessibility	Reference to instant access and ease of use	27
Student preferences	Preference for combined human and AI feedback	24

Note. Frequencies reflect how often each theme appeared in student survey responses.

The thematic analysis of student survey responses and the instructor’s reflective memo revealed four key themes: (1) clarity and specificity, (2) timeliness and accessibility, (3) student preferences, and (4) instructor reflections.

Clarity and specificity

Students generally appreciated the precision with which ChatGPT identified grammatical and lexical issues. For example, one participant described the AI-generated feedback as follows:

It showed me many grammar [grammar] mistakes very clearly and help [helped] me correct my sentence faster. It also gave me better words and fixed small language mistakes that I did not notice before. Sometimes when I write long sentences, ChatGPT explains which part is not correct and how I can write it in more clear and academic way. This makes my writing look more professional and easy [easier] to understand.

However, many students reported that the feedback was somewhat formulaic and lacked the depth needed to address higher-order concerns such as coherence, tone, and argument development. In contrast, students noted that the instructor's feedback was more contextualized and helped them consider how to reshape their ideas. As one participant noted:

Sometimes ChatGPT feedback was too general and repeated similar comments many times. It told me to improve my organization and develop my ideas more, but I was not always understanding exactly what I should change in my essay. My teacher feedback was more helpful because the teacher explain[ed] clearly why some examples were not connected strongly to my argument and how I can [could] organize my ideas in [a] more logical way.

The instructor's feedback often addressed rhetorical choices and the development of the argument, which many students found especially useful. These findings align with observations by Gayed et al. (2022) and Farzi (2024), who emphasized the gap in rhetorical scaffolding when feedback is generated solely by AI.

Timeliness and accessibility

Students valued ChatGPT's immediate feedback, which enabled quicker revision cycles and encouraged iterative writing practices. Although instructor feedback was delayed, students perceived it as more targeted and responsive to individual learning needs. These findings suggest that AI can support time-sensitive revision in large classroom settings, a conclusion also reached by Steiss et al. (2024).

Student preferences

Most students preferred receiving both types of feedback. They viewed ChatGPT as useful for technical corrections but relied on the instructor for in-depth, actionable guidance. A few students described AI feedback as overly mechanical.

As one participant explained:

I prefer both teacher and ChatGPT feedback because they help me in different ways. ChatGPT was useful for grammar correction and vocabulary, but the teacher feedback explained my mistakes more deeply and help[ed] me improve my ideas and essay organization better.

These perceptions are consistent with research on students' preferences for feedback modalities. Escalante et al. (2023) found that some students preferred AI-

generated feedback for its clarity and specificity, whereas others valued human feedback for its contextual sensitivity; many reported that combining both forms in a blended feedback environment offered the greatest benefit across dimensions of writing improvement.

Instructor reflections

The instructor noted that ChatGPT's analysis of grammar and surface-level features was largely accurate and helpful to students. However, they also observed limitations in its handling of rhetorical nuance, particularly when student writing required cultural or contextual interpretation. The formulaic style of the AI-generated comments sometimes made the feedback seem overly structured and repetitive. Similarly, Farzi (2024) identified limitations in AI output when evaluating writing involving complex discursive elements.

Combined interpretation

The triangulated findings from quantitative score comparisons, student survey responses, and instructor reflections indicate that ChatGPT-3.5 consistently identifies surface-level features and assigns broadly comparable scores. Its limitations were more apparent, however, in evaluations of rhetorical complexity and context-dependent writing. In these areas, instructor feedback remained essential to supporting deeper revision and meaningful improvement. The findings support a combined approach in which the speed and reach of AI tools are complemented by instructors' pedagogical judgment. As Farzi (2024) argues, calibrated generative AI used within a clear, guided feedback process may improve efficiency without lowering academic standards. This integration may be particularly useful in larger classes, in which instructors often have limited time to provide personalized feedback to every student.

Discussion

The findings have important implications for integrating generative AI into EAP writing instruction. Although the study provided evidence of potential usefulness of ChatGPT-3.5 for providing timely and consistent surface-level feedback, it also revealed limitations in addressing higher-order writing skills that require contextual awareness, rhetorical sensitivity, and pedagogical insight.

The alignment between human and AI scoring suggests that ChatGPT-3.5 can assess grammar and other surface-level features with a reasonable level of reliability. The small difference in mean scores and nonsignificant p-value support its potential as a supplementary rater in formative assessment. The boxplot (Figure 1) provides a descriptive visualization of this similarity as the distribution of ChatGPT-3.5 scores resembles the pattern seen in the human ratings.

At the same time, ChatGPT's slightly lower mean score may reflect a more rigid application of the rubric. The model appears to apply the criteria strictly, with limited adjustment for context or student intention. In contrast, human raters can draw on teaching experience to interpret what a student is attempting to achieve, a capacity that GenAI does not fully replicate.

The qualitative data revealed nuanced student preferences and expectations regarding writing feedback. Students valued the speed and clarity of AI-generated suggestions but trusted human feedback more because it was more detailed, captured the writing context more effectively, and helped them reshape their ideas. These findings align with previous studies (Gayed et al., 2022; Steiss et al., 2024) and underscore the continuing importance of human feedback in learning.

For example, the theme of “clarity and specificity” indicates that, although GenAI provided effective grammar corrections, it offered limited insight into coherence and tone. Similarly, under “timeliness and accessibility,” students appreciated the rapid feedback cycle but continued to rely on human feedback for deeper revision. These complementary patterns support the practicality of a hybrid model in which GenAI provides preliminary feedback on drafts and human feedback guides refinement (see Farzi, 2024).

This study offers important insights for educators exploring ways to integrate GenAI into academic writing instruction. It highlights the potential benefits and necessary conditions for using GenAI tools effectively, especially in L2 learning environments.

The findings indicate that generative AI works best as a supplement to instructor feedback, not as a replacement. In L2 writing instruction, tools such as ChatGPT-3.5 can identify recurring surface-level issues, allowing instructors to focus on higher-order writing strategies. A growing body of research supports this view. For example, Karagoz (2025) found that, although AI tools effectively provide surface-level corrections, they lack the contextual sensitivity required for deeper content- and discourse-level feedback. Similarly, Luo et al. (2025) observed that blended feedback models combining AI suggestions with instructor input led to greater writing improvement among L2 learners.

As Farzi (2024) argues, a structured approach involving rubric calibration, prompt design, and iterative feedback can mitigate some of GenAI’s limitations and strengthen its classroom value. Emerging AI-supported learning environments may allow instructors to draw more directly on course rubrics, model texts, and other instructional resources when generating feedback. This approach may produce responses that align more closely with classroom expectations and instructional priorities than feedback generated through isolated prompts (Fan et al., 2024; Lewis et al., 2021).

This approach could also support more equitable access to feedback in large classes and encourage peer-review models in which students engage with both AI- and peer-generated feedback before receiving instructor input. However, its effectiveness depends on students’ ability to evaluate feedback critically. Li (2025) emphasized that critical AI literacy is essential to helping students evaluate the relevance, reliability, and ethical implications of automated feedback rather than accept AI-generated responses uncritically.

Beyond L2 teaching, this study contributes to discussions of responsible AI use in education. As K. V. Nguyen (2025) highlights, ethical AI use requires an understanding of how algorithms operate, attention to fairness, and protection of user

data. The findings suggest that AI may support some teaching tasks, but its use should prioritize student learning rather than technology adoption for its own sake.

The impact of AI feedback depends on how it is introduced and explained. Instructors should teach writing while also helping students evaluate the different forms of feedback they receive. This approach supports the view advanced by Farzi (2024) of the instructor's changing role as a mediator between AI systems and human judgment. Tools such as ChatGPT may be more effective when used within a clear instructional strategy that includes calibrated prompts and deliberate planning.

Limitations

Although this study offers insight into the potential of ChatGPT-3.5 to support formative feedback in EAP writing instruction, several limitations should be considered.

First, the relatively small sample ($N = 30$) limits the generalizability of the findings. The sample was appropriate for an exploratory study; however, a larger and more diverse group could yield more robust evidence and better represent learners from different linguistic backgrounds, academic disciplines, and levels of English proficiency.

Second, the writing tasks were limited to only one genre: argumentative writing. Academic writing encompasses many other genres, such as expository, discursive, cause-and-effect, problem-and-solution, reflective, and reports. Because each task has distinct structural and stylistic conventions, GenAI may respond differently to each. Future studies should include additional genres to clarify how GenAI feedback functions across types of academic writing.

Third, ChatGPT-3.5 was used because it was the version available when the data were collected and analyzed. Although the model was calibrated across three stages, it is unclear whether the findings would apply to newer models or other GenAI tools, such as GPT-4 or GPT-5 (OpenAI, 2025). As these technologies develop, research should examine how their performance changes, whether similar calibration remains necessary, and how appropriately they can support formative and summative feedback in L2 writing instruction. Comparisons across models and platforms could provide a better understanding of which tools are most useful for academic writing support.

Fourth, the instructor's involvement centred primarily on rubric design, essay scoring, and written feedback. The study did not examine teacher-student interactions around AI-generated feedback during the writing process. Examining these real-time interactions could provide deeper insight into how students, instructors, and GenAI tools work together to support classroom learning.

Finally, ethical concerns require further attention. The use of GenAI in writing classrooms raises questions about system transparency, the handling of student data, and potential bias in feedback. Students' self-reported perceptions may also have been shaped by their previous experiences with or attitudes toward technology. Future research should incorporate objective performance measures, triangulate data sources, and follow clear ethical principles and policies. Such measures could help ensure that GenAI is used responsibly and equitably to support student learning.

Conclusion

This study examined how ChatGPT-3.5 can support formative feedback in EAP writing instruction when used under careful instructor guidance. The quantitative findings showed a strong correlation between ChatGPT-3.5 scores after the third calibration phase and the human instructor's scores. Qualitatively, students described AI-generated feedback as immediate and useful for technical issues such as grammar and vocabulary. At the same time, they strongly preferred instructor feedback for its greater depth and contextual sensitivity.

These findings suggest that combining AI-generated feedback with instructor input may improve feedback quality, reduce instructor workload, and support student progress. Such a model may be particularly useful in large or writing-intensive courses, in which instructors may have limited time to provide individualized feedback.

For AI tools to be effective in writing instruction, they must be situated within a clear pedagogical framework. Students need guidance in evaluating feedback, identifying its source, and deciding how to use it in their writing. Instructors remain central to this process. The role of GenAI in the L2 classroom is not to replace instructors but to support pedagogy, improve instructional efficiency, and strengthen student learning.

Further research should examine the longitudinal effects of GenAI-supported feedback on student performance, its effectiveness across different writing genres, and its use in diverse academic contexts. As GenAI tools evolve, research and pedagogy must evolve with them.

Acknowledgments

I thank the EAP students who participated in this study for their time, engagement, and willingness to contribute. Their participation made this research possible.

References

- Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, 7(1), 54–74.
<https://doi.org/10.1080/15434300903464418>
- Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198.
<https://doi.org/10.18653/v1/2020.acl-main.463>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.
<https://doi.org/10.1191/1478088706qp063oa>
- Cotos, E. (2011). Potential of automated writing evaluation feedback. *CALICO Journal*, 28(2), 420–459. <https://doi.org/10.11139/cj.28.2.420-459>
- Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research* (3rd ed.). SAGE.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv: [1810.04805](https://arxiv.org/abs/1810.04805).
arXiv. <https://arxiv.org/abs/1810.04805>

- Dikli, S. (2006). Automated essay scoring. *Turkish Online Journal of Distance Education*, 7(1), 49–62. <https://dergipark.org.tr/en/pub/tojde/article/176620>
- 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, Article 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>
- Fan, Y., Tan, S., & Lim, G. Y. W. (2024). EAP teacher feedback in the age of AI: Supporting first-year students in EFL disciplinary writing. *Australian Journal of Applied Linguistics*, 7(3), 250–267. <https://doi.org/10.29140/ajal.v7n3.1943>
- Farzi, R. (2024). Calibrating generative AI for second language writing assessment: Combining statistical validation with prompt design. *Assessment and Practice in Educational Sciences*, 2(4), 1–12. <https://journalapes.com/index.php/apes/article/view/91>
- Ferris, D. R. (2014). Responding to student writing: Teachers' philosophies and practices. *Assessing Writing*, 19, 6–23. <https://doi.org/10.1016/j.asw.2013.09.004>
- Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence*, 6, Article 1162454. <https://doi.org/10.3389/frai.2023.1162454>
- García-López, I. M., & Trujillo-Liñán, L. (2025). Ethical and regulatory challenges of generative AI in education: A systematic review. *Frontiers in Education*, 10, Article 1565938. <https://doi.org/10.3389/feduc.2025.1565938>
- Gayed, J. M., Carlon, M. K. J., Oriola, A. M., & Cross, J. S. (2022). Exploring an AI-based writing assistant's impact on english language learners. *Computers and Education: Artificial Intelligence*, 3, Article 100055. <https://doi.org/10.1016/j.caeai.2022.100055>
- Harunasari, S. Y. (2023). Examining the effectiveness of AI-integrated approach in EFL writing: A case of ChatGPT. *International Journal of Progressive Sciences and Technologies*, 39(2), 357–368. <https://doi.org/10.52155/ijpsat.v39.2.5516>
- Hyland, K. (2016). *Teaching and researching writing* (3rd ed.). Routledge.
- Hyland, K., & Hyland, F. (Eds.). (2019). *Feedback in second language writing: Contexts and issues* (2nd ed.). Cambridge University Press. <https://doi.org/10.1017/9781108635547>
- Kaliisa, R., Misiejuk, K., López-Pernas, S., & Saqr, M. (2026). How does artificial intelligence compare to human feedback? a meta-analysis of performance, feedback perception, and learning dispositions. *Educational Psychology*, 46(1), 80–111. <https://doi.org/10.1080/01443410.2025.2553639>
- Karagoz, I. (2025). AI-generated feedback in english writing instruction for language learners: A systematic review. *The Reading Matrix: An International Online Journal*, 25(1), 51–67. <https://www.readingmatrix.com/files/34-9gp90gjo.pdf>
- Landauer, T. K. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice*, 10(3), 295–308. <https://doi.org/10.1080/0969594032000148154>
- Lee, S. S., & Moore, R. L. (2024). Harnessing generative AI (GenAI) for automated feedback in higher education: A systematic review. *Online Learning*, 28(3), 82–104. <https://doi.org/10.24059/olj.v28i3.4593>

- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). *Retrieval-augmented generation for knowledge-intensive NLP tasks*. arXiv. <https://doi.org/10.48550/arXiv.2005.11401>
- Li, S. (2025). Generative AI and second language writing. *Digital Studies in Language and Literature*, 2(1), 122–152. <https://doi.org/10.1515/dsl-2025-0007>
- Liang, W., Zhang, Y., Codreanu, M., Wang, J., Cao, H., & Zou, J. (2025). *The widespread adoption of large language model-assisted writing across society*. arXiv: 2502.09747. arXiv. <https://arxiv.org/abs/2502.09747>
- Lund, B. D., Lee, T. H., Mannuru, N. R., & Arutla, N. (2025). AI and academic integrity: Exploring student perceptions and implications for higher education. *Journal of Academic Ethics*, 23(3), 1545–1565. <https://doi.org/10.1007/s10805-025-09613-3>
- Luo, M., Hu, X., & Zhong, C. (2025). The collaboration of AI and teacher in feedback provision and its impact on EFL learners' argumentative writing. *Education and Information Technologies*, 30, 17695–17715. <https://doi.org/10.1007/s10639-025-13488-7>
- McNamara, T. (2000). *Language testing*. Oxford University Press.
- Nelson, A. S., Santamaría, P. V., Javens, J. S., & Ricaurte, M. (2025). Students' perceptions of generative artificial intelligence (GenAI) use in academic writing in english as a foreign language. *Education Sciences*, 15(5), Article 611. <https://doi.org/10.3390/educsci15050611>
- Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human-AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education*, 49(5), 847–864. <https://doi.org/10.1080/03075079.2024.2323593>
- Nguyen, K. V. (2025). The use of generative AI tools in higher education: Ethical and pedagogical principles. *Journal of Academic Ethics*, 23, 1435–1455. <https://doi.org/10.1007/s10805-025-09607-1>
- OECD. (2026). *OECD digital education outlook 2026: Exploring effective uses of generative AI in education*. <https://doi.org/10.1787/062a7394-en>
- Ogunleye, B., Zakariyyah, K. I., Ajao, O., Olayinka, O., & Sharma, H. (2024). A systematic review of generative AI for teaching and learning practice. *Education Sciences*, 14(6), Article 636. <https://doi.org/10.3390/educsci14060636>
- OpenAI. (2023). ChatGPT (GPT-3.5). <https://chatgpt.com/>
- OpenAI. (2025). ChatGPT (GPT-5). <https://chatgpt.com/>
- Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Lawrence Erlbaum Associates.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- Rana, A., Vaidya, P., & Hu, Y. C. (2026). Transformative role of large language models in education and comparative analysis with traditional technologies. *Education and Information Technologies*, 31(4), 1137–1162. <https://doi.org/10.1007/s10639-025-13854-5>
- Samala, A. D., Rawas, S., Wang, T., Reed, J. M., Kim, J., Howard, N.-J., & Ertz, M. (2024). Unveiling the landscape of generative artificial intelligence in education: A comprehensive taxonomy of applications, challenges, and future prospects.

- Education and Information Technologies*, 30, 3239–3278.
<https://doi.org/10.1007/s10639-024-12936-0>
- 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., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback on students' writing. *Learning and Instruction*, 91, Article 101894.
<https://doi.org/10.1016/j.learninstruc.2024.101894>
- Uludag, P., & McDonough, K. (2022). Validating a rubric for assessing integrated writing in an EAP context. *Assessing Writing*, 52, Article 100609.
<https://doi.org/10.1016/j.asw.2022.100609>
- van Niekerk, J., Delport, P. M., & Sutherland, I. (2025). Addressing the use of generative AI in academic writing. *Computers and Education: Artificial Intelligence*, 8, Article 100342. <https://doi.org/10.1016/j.caeai.2024.100342>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
<https://proceedings.neurips.cc/paper/7181-attention-is-all-you-need>
- Wang, D. (2024). Teacher- versus AI-generated (Poe application) corrective feedback and language learners' writing anxiety, complexity, fluency, and accuracy. *The International Review of Research in Open and Distributed Learning*, 25(3), 37–56. <https://doi.org/10.19173/irrodl.v25i3.7646>
- Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, 11(1), Article 28. <https://doi.org/10.1186/s40561-024-00316-7>
- Zhang, Z., Aubrey, S., Huang, X., & Chiu, T. K. F. (2025). The role of generative AI and hybrid feedback in improving L2 writing skills: A comparative study. *Innovation in Language Learning and Teaching*. Advance online publication.
<https://doi.org/10.1080/17501229.2025.2503890>

Appendix A: EAP writing assessment rubric

Criteria	Poor (1)	Needs improvement (2)	Somewhat meets expectations (3)	Meets expectations (4)	Exceeds expectations (5)
Lexical accuracy and range	Uses a basic range of vocabulary, including some academic terms. Errors in word form and word choice sometimes interfere with meaning. Limited control over collocations.	Uses a developing range of academic and general vocabulary. Errors in word form and word choice may sometimes interfere with meaning. Collocations are occasionally inaccurate.	Uses an adequate range of academic and abstract vocabulary but may lack precision in specialized terms. Errors occasionally obscure meaning. Collocations are often correct.	Uses an effective range of academic and figurative vocabulary. Errors in word form and word choice are infrequent and minor, but precision in complex ideas may still be affected. Collocations are mostly accurate, though occasional awkward phrasing may appear.	Uses an advanced range of academic and figurative vocabulary. Word choice and collocations are almost always precise, even in complex or abstract contexts. Errors are very rare and do not impact meaning.
Grammatical accuracy and range	Uses basic sentence structures accurately and attempts some complex forms with partial success. Frequent errors in verb tense and sentence structure sometimes interfere with meaning.	Uses a mix of simple and complex structures effectively, with growing control over verb tenses and auxiliaries. Errors in advanced grammar may sometimes interfere with meaning.	Uses a range of simple and complex structures, but errors in advanced forms occasionally obscure meaning. Verb tense control is improving but remains inconsistent.	Uses a broad range of simple and complex structures. Errors are infrequent and minor and do not significantly interfere with meaning. Verb tense and auxiliary control are generally accurate.	Uses a wide variety of complex sentence structures effectively. Errors, if present, are very rare and do not impact meaning. Verb tense and auxiliary control are highly accurate.
Organization and development of the topic	Presents ideas in a basic, loosely structured manner with limited logical flow. Uses simple transitions, but connections between ideas are weak. Ideas are minimally developed, often repetitive or unclear. Paraphrasing is often inaccurate or overly reliant on original wording.	Organizes writing into clear sections, though connections between ideas may still be weak. Uses a wider range of transitions, though inconsistently. Ideas show moderate development, with some explanation and support. Errors in paraphrasing or referencing sometimes obscure meaning.	Demonstrates clear structure with improving coherence. Uses transitions effectively, though they may still be formulaic. Ideas are developed, but support and explanation may lack depth. Sources are integrated with improving accuracy, though occasional paraphrasing errors may affect clarity.	Demonstrates strong logical organization, with a clear and coherent flow. Uses varied and natural transitions. Ideas are well-developed, with effective support. Sources are integrated effectively, with paraphrasing that maintains original meaning.	Demonstrates clear and effective organization, ensuring a logical flow of ideas. Uses natural, well-placed transitions. Ideas are fully developed, with well-supported explanations. Synthesizes sources confidently, integrating accurate paraphrasing.
Overall effectiveness	Communicates basic meaning but with frequent awkward phrasing. Extra effort is required from the reader to understand the message.	Communicates meaning with improving fluency. Occasional awkward phrasing or errors may sometimes interfere with clarity, requiring some extra effort from the reader.	Communicates meaning effectively for most topics. Awkward phrasing or unclear transitions occasionally obscure meaning. Minimal extra effort is required from the reader.	Communicates meaning with clarity and fluency. Writing generally sounds natural and requires very little extra effort from the reader.	Writing is clear, fluent, and well-structured. Writing reads naturally and requires no extra effort from the reader.

Appendix B: Prompts for the three stages of the EAP writing assessment study

Baseline AI scoring prompt (Stage 1)

Hey ChatGPT, imagine you are an evaluator using the EAP Writing Assessment Rubric. Read the essay carefully and assign a score for each category. Provide a brief justification for each score with reference to the rubric and give specific suggestions for improvement. Finally, assign an overall grade based on the rubric criteria.

Clarified-descriptor AI scoring prompt (Stage 2)

For our second round, please act as an evaluator using the EAP Writing Assessment Rubric. You were too stringent in the first round and your scores were lower than our expert evaluator. When scoring, balance grammatical and lexical accuracy with discourse-level qualities such as thesis clarity, paragraph structure, logical progression, and argumentation. Give equal weight to discourse-level qualities and linguistic accuracy. Minor language errors that do not impede meaning should not result in significant penalties. Provide category scores with justifications that reference both strengths and areas for improvement and offer improvement suggestions that address both higher-order and lower-order issues. Finally, assign an overall grade based on the rubric criteria.

Final calibrated AI scoring prompt with performance anchors (Stage 3)

For our third stage, please act as an evaluator using the EAP Writing Assessment Rubric.

Performance anchors are representative examples of writing at different quality levels that serve as benchmarks for scoring. They show what high, mid, and low performance look like in practice and help evaluators apply the rubric more consistently. You will use these examples as concrete reference points when assigning scores.

You now have excerpts for high, mid, and low performance for each rubric category. Use them to compare the target essay and to guide your scoring, just as human raters do during training. Consider both strengths and weaknesses in each category, ensuring that no single aspect such as grammatical accuracy dominates your evaluation. Minor language errors that do not impede communication should have minimal impact on scores.

High performance example

Excerpt: The increasing reliance on renewable energy is not merely an environmental necessity; it is an economic opportunity. Countries that invest in solar and wind infrastructure now will secure long-term energy independence and job growth. For instance, Germany's renewable sector employs over 300,000 workers, demonstrating that environmental policy and economic prosperity can coexist.

Our expert evaluator scores:

- Lexical accuracy and range: 5
- Grammatical accuracy and range: 5
- Organization and development: 5
- Overall effectiveness: 5

Mid performance example

Excerpt: Renewable energy is important because it can help environment and create the jobs. For example, some countries make solar power and also wind power. This help economy grow. But sometimes the cost is high and some people do not like it. In the future, more research is necessary to make it better.

Our expert evaluator scores:

- Lexical accuracy and range: 3.5
- Grammatical accuracy and range: 3
- Organization and development: 3.5
- Overall effectiveness: 3.5

Low performance example

Excerpt: Renew energy good, people use sun and wind. Some country have this but not all. This help planet but money problem. Jobs maybe come but not sure. Government need to do something.

Our expert evaluator scores:

- Lexical accuracy and range: 1
- Grammatical accuracy and range: 2
- Organization and development: 1
- Overall effectiveness: 1

Instructions for Evaluation

Read the target essay carefully.

Assign a score for each category using the rubric and the performance anchors as benchmarks. Justify each score with explicit reference to both the rubric and the examples. Provide prioritized improvement suggestions that address both higher order discourse issues and lower order language issues. Assign an overall grade based on the rubric criteria.

Appendix C: Student perceptions of AI use survey

To protect your privacy and keep your responses anonymous, enter the first two letters of your first name, followed by the last two digits of your phone number, and then the first two letters of your mother's name. For example: HO23FA. _____

Closed-ended questions (1–12)

1. How often did you use AI during this course?
☐ Daily ☐ A few times per week ☐ Weekly ☐ Occasionally ☐ Never
2. Which AI tools did you use most this term? (Check all that apply.)
☐ ChatGPT ☐ Gemini ☐ Copilot ☐ Grammarly AI ☐ Other: _____
3. What language skills did you use AI to improve? (Check all that apply.)
☐ Writing ☐ Speaking ☐ Reading ☐ Listening ☐ Vocabulary ☐ Other: _____
4. Did AI help you improve your writing?
☐ Yes ☐ No ☐ Not sure
5. Did AI help you improve your speaking?
☐ Yes ☐ No ☐ Not sure
6. Did AI help you improve your vocabulary?
☐ Yes ☐ No ☐ Not sure
7. Did AI help you improve your reading or listening comprehension?
☐ Yes ☐ No ☐ Not sure
8. How useful was AI-generated feedback by the end of the course?
☐ Not useful ☐ Slightly useful ☐ Moderately useful ☐ Very useful
☐ Extremely useful
9. Compared to instructor feedback, how helpful was AI feedback?
☐ More helpful ☐ About the same ☐ Less helpful ☐ I didn't receive AI feedback
10. Did AI help you feel more confident in your English skills?
☐ Yes ☐ No ☐ Not sure
11. Would you like AI tools to be used more in future ESL courses?
☐ Yes ☐ No ☐ Not sure
12. Would you recommend using AI tools to other ESL students?
☐ Yes ☐ No ☐ It depends

Open-ended questions (13–16)

13. How clear and specific was the feedback you received from AI? Please describe whether it helped with grammar, vocabulary, or other areas of your writing.
14. What did you think about the speed, accessibility, or ease of using AI feedback during the course?
15. When it comes to feedback on your work, which do you prefer: instructor feedback, AI feedback, or a combination of both? Please explain why.
16. What suggestions do you have for improving the use of AI feedback in future ESL classes?