OPEN ACCESS

Australian Journal of Applied Linguistics
ISSN 2209-0959
https://www.castledown.com/journals/ajal/

*Australian Journal of Applied Linguistics*, 9, 103200 (2026)
https://doi.org/10.29140/ajal.2026.103200

# *Two voices, one draft: a diachronic comparison of human and AI feedback in EFL essay writing*

SZU-YU RUBY CHEN

*Chung Yuan Christian University, Taiwan*

## Abstract

As intelligent technological tools increasingly permeate educational contexts, chatbots have emerged as prominent aids in language learning. This study investigates the comparative effectiveness of human- and AI-generated feedback on EFL learners' essay writing, focusing on differences and similarities in feedback focus, specificity, and rubric alignment. Specifically, the study examines how feedback from a human instructor and a generative AI (GenAI) system, GPT-4, varies across three dimensions: structure and organisation, content and ideas, and grammar. A diachronic perspective is used to analyse students' observed revision development across drafts under instructor feedback, alongside a post hoc comparison of GPT-4 feedback applied to the same baseline texts.

Participants were twenty-six EFL college students enrolled in an English composition course. The study focused on the production of comparison essays through a three-stage process. In the first stage, students independently wrote their initial drafts at home, with access to dictionaries or online resources but without direct assistance. In the second stage, the drafts received written feedback from a human instructor. In the third stage, GPT-4 generated feedback on students' initial drafts retrospectively for analytic comparison only; students did not receive or use AI feedback during revision. Instructor scores and GPT-4 scores were derived using the same rubric categories to enable controlled comparison. Findings show partial alignment between GPT-4 and instructor feedback, particularly in identifying recurring local issues such as grammatical accuracy and sentence-level clarity. However, qualitative comparisons indicate clearer divergence on higher-order concerns. For example, instructor feedback provided richer, context-sensitive guidance related to task fulfilment, rhetorical appropriacy, and content elaboration, whereas GPT-4 feedback was more standardised and rubric-driven, often foregrounding surface-level edits and general recommendations.

These results underscore the potential value of integrating GenAI tools into EFL writing instruction while also highlighting current limitations. Pedagogically, the findings support blended feedback models in which AI supplements rapid local editing and initial diagnostic support, while teachers provide meaning-level guidance and mediate feedback uptake. The study contributes to a deeper understanding of how human and AI feedback may complement each other and offers pedagogical insights into the thoughtful use of AI in academic writing contexts.

*Keywords:* EFL writing, human-/AI- feedback, diachronic revision analysis

## Introduction

### *Background of the Research*

In recent decades, there has been a growing interest in the integration of technology, ushering in a new era in the field of pedagogy. Web-based platforms, in particular, have gained widespread popularity as valuable tools in language teaching and learning. The increasing adoption of web-based platforms by researchers in classrooms raises an important issue for educators, i.e. the effective incorporation of technological tools into language pedagogy. As technology becomes an integral part of educational settings, the use of chatbots as a learning assistant is garnering increased attention. This extends to various contexts, including first language acquisition, second language learning, and foreign language education. Chatbots, particularly those powered by generative artificial intelligence (GenAI), now function not only as conversational tools but also as semi-autonomous instructional agents capable of offering writing feedback.

Previous studies on chatbots have primarily focused on information delivery and responding to frequently asked questions through rule-based or retrieval systems (Smutny and Schreiberova, 2020). In educational settings, chatbots have traditionally played in supporting students' adaptation to university life (Carayannopoulos, 2018) and aiding instructors in managing large in-class activities (Schmulian and Coetzee, 2019). Most existing studies have examined chatbots in the context of class engagement (Kerly et al., 2007) or the development of conversational competence (Fryer & Carpenter, 2006). Far fewer have explored how generative AI can support the complex cognitive and linguistic processes involved in writing. Writing, especially in the context of English for Academic Purposes (EAP), requires more than grammatical accuracy; it demands lexical richness, cohesion, and the ability to develop structured arguments (Kroll, 2001).

With the rapid advancement of generative AI models such as ChatGPT, the role of chatbots in education has shifted from simple task automation to more dynamic, language-based interaction. Numerous studies have highlighted the positive impact of web-based platforms on students' language learning. However, despite the growing capabilities of AI-powered chatbots, there remains a relative scarcity of research focusing on the qualitative nature and comparative impact of feedback in AI-assisted writing instruction. Recent empirical studies have begun to address this gap, offering nuanced insights into the characteristics, functions, and pedagogical value of feedback provided by both human instructors and AI chatbots in academic writing contexts (Zou et al., 2025; Chen et al., 2024; Escalante et al., 2023). Although models like ChatGPT have demonstrated efficacy in addressing grammar and organisation (Klipp & Reinders, 2025), learners frequently hesitate to adopt AI-generated feedback at the level

of content development or argumentation (Chen et al., 2024). These limitations underscore the need for explicit pedagogical scaffolding and improved feedback literacy in technology-enhanced writing instruction.

Chatbots can function as tireless assistants, relieving instructors of repetitive tasks and offering learners consistent, on-demand language support (Fryer et al., 2019). However, recent findings reveal that learners frequently reject or disregard AI-generated feedback that lacks specificity, misinterprets their intentions, or overwhelms them with vague suggestions (Chen et al., 2024). This study explores how the complementary strengths of chatbot-generated feedback and human-provided feedback can be leveraged in EFL writing classrooms. For instance, in English writing courses, where instructional goals extend beyond grammar reinforcement to the development of logical reasoning and critical writing skills, feedback plays a crucial role. Drawing from recent teaching experiences, the instructor has observed that many students enrolled in composition courses encounter persistent challenges, particularly in generating well-developed content in English writing, with a significant proportion of these learners demonstrating low performance in this area.

To address these challenges, this study traces students' draft-to-draft revision development under instructor feedback and conducts a retrospective, draft-level comparison between instructor feedback and GPT-4 feedback applied to the same initial texts; AI feedback was not available to students and is analysed for comparative purposes only. The objective is to identify qualitative differences between the two feedback sources, examining their feedback focus, specificity, and rubric alignment, rather than their relative "effectiveness" in shaping students' revisions. This diachronic perspective derives from the draft-to-draft revision trajectory under instructor feedback and is used to clarify how teacher mediation functions as an instructional mechanism across successive revisions. Post hoc GPT-4 feedback serves as a comparative benchmark of what the AI system would likely foreground on the same baseline texts. While previous studies have largely concentrated on surface-level features such as grammar and syntax, little attention has been paid to how AI tools address meaning-level writing attributes such as argument coherence, idea elaboration, and audience awareness. Recent longitudinal studies suggest that, when implemented strategically, AI-generated feedback can match human input in promoting writing improvement (Escalante et al., 2023), although learners' engagement with and uptake of automated feedback vary considerably (Zou et al., 2025).

Thus, this study lies at the intersection of language education and educational technology, comparing AI-generated feedback with that from human instructors in terms of content, clarity, and instructional depth. It further investigates whether AI feedback can not only support higher-order writing skills but also foster learner autonomy in revision and self-directed learning. This focus aligns with current pedagogical recommendations for blended feedback models that integrate the efficiency of AI with the contextual richness of human instruction (Dizon et al., 2025; Klipp & Reinders, 2025).

### *Research Questions*

At the beginning of the study, the researcher underscores the application of ChatGPT in EFL writing, with a focus on both its affordances and limitations as a feedback tool. The research seeks to investigate not only the social and instructional benefits of chatbots in writing instruction but also how AI feedback differs qualitatively from human feedback in areas such as depth, personalisation, and responsiveness to meaning-level writing concerns. While learners often use ChatGPT to address surface-level issues such as grammar and vocabulary (Klipp & Reinders, 2025), they may find it challenging to apply AI feedback effectively in more cognitively

demanding writing tasks without appropriate guidance (Chen et al., 2024; Dizon et al., 2025) The central research question guiding this study is as follows:

(1) What qualitative differences exist between instructor feedback and post hoc GPT-4 feedback on students' essay drafts?
(2) To what extent do students' instructor-guided revisions align with the issues highlighted by post hoc GPT-4 feedback?

By comparing the characteristics of both human and AI feedback, this study aims to illuminate the potential for integrating both feedback types to improve EFL writing instruction. The findings of this study contribute to the growing body of research on chatbot integration in language education, particularly in the context of EFL writing. Recent research encourages educators to take an active role in helping learners engage with AI feedback critically and ethically, integrating it as a scaffolded tool rather than a replacement (Dizon et al., 2025; Escalante et al., 2023). The study aims to propose a pedagogical framework that considers the qualitative distinctions between AI and human feedback and how they might best complement one another.

In summary, the structure of the research runs as follows, as illustrated in Figure 1 (Napkin, 2025):

(1) To examine and reconceptualise feedback practices in EFL writing instruction through a diachronic comparison of AI-generated and human-provided feedback.
(2) To explore how chatbot technologies can be integrated into EFL pedagogy to enhance learners' cognitive engagement, foster critical thinking, and support the development of coherent, content-rich academic writing in the future EFL writing classroom.

In the following section, theoretical concepts relevant to this investigation are discussed, including: (2.1) pedagogical potentials of chatbots; (2.2) web-based and AI-supported platforms in EFL writing development; and (2.3) influence of human and AI feedback in EFL writing instruction.

## Literature Review

### *Pedagogical Potentials of Chatbots*

A chatbot is a software-based conversational agent designed to engage users in natural-language interaction via text or voice within a bounded domain. Chatbots have been implemented


**Figure 1.** *Aim of the study. Created using Napkin (Version 2025)*

across multiple sectors, including customer service, technical support, education and training (Smutny & Schreiberova, 2020), and are now discussed within broader international work on technology-mediated learning environments, including research on cloud-based smart technologies for open learning (Papadakis et al., 2023a, 2023b) and cross-disciplinary syntheses on the educational value of AI-enabled agents and social robots (Lampropoulos & Papadakis, 2025). As digital infrastructures mature, interaction with conversational systems has become a routine means of accessing information, sustaining dialogue, and completing tasks. Text-based chatbots have historically operated through predefined rules or algorithms that map user inputs to system outputs (Budiu, 2018). The development of pedagogical agents has further encouraged educators to adopt messaging-based tools to support instruction, including systems such as Ethnobot, which simulates ethnographic interviewing to elicit qualitative data through structured conversational prompts. Chatbots have also been positioned as potentially valuable in online and distance education contexts (Heller et al., 2005) and, more recently, large-language-model systems have re-centred scholarly attention on the quality, prioritisation, and actionability of machine-generated formative comments in writing tasks (e.g., Steiss et al., 2024).

However, the literature consistently indicates that "chatbots" are not pedagogically uniform; rule-based systems and contemporary AI-driven models differ in interactional capacity, feedback granularity, and associated epistemic risks such as unwarranted confidence in system outputs. Accordingly, claims regarding pedagogical benefit require explicit specification of system type, instructional function, such as practice partner vs. feedback provider, and the learning mechanisms targeted by the task design. Without such specification, findings across studies are difficult to compare, and pedagogical claims risk overgeneralisation.

Research on technology-mediated writing and literacy development further suggests that performance gains are rarely attributable to "technology" per se. Digital tools have been associated with improved writing outcomes, because they reduce procedural barriers and enable more efficient engagement with language tasks (Peregoy & Boyle, 2012). For example, Alsaleem (2014) reported improvements in learners' writing and speaking through WhatsApp-based dialogue journals, while Lin and Yang (2011) found that wiki-based collaborative writing supported development through timely peer and platform-mediated feedback. Taken together, these studies underscore that writing gains are mediated by pedagogical design, particularly task structure and the surrounding feedback ecology. Learning benefits are most evident when tools externalise revision trajectories, expand opportunities for interaction, and scaffold learners' uptake and use of feedback. Accordingly, this literature shifts the analytic emphasis from whether digital tools "work" to the conditions under which they reorganise interaction and feedback in ways that generate sustained, developmentally meaningful revision.

Within this broader landscape, chatbots have been framed as a distinctive class of tools for language learning because they can provide on-demand interaction, diversified input, and rapid feedback (Fryer et al., 2019; Jia et al., 2012). Relative to human peers, chatbots may expand learners' linguistic exposure during drafting. Belda-Medina and Calvo-Ferrer (2022), for instance, suggested that chatbots can introduce varied input that supports idea elaboration and language noticing. Chatbots may also lower affective barriers by offering a low-stakes environment in which learners can experiment with language without fear of negative evaluation (Jia & Ruan, 2008). Yet this affective advantage is not necessarily developmental. Constant availability and non-judgmental interaction may promote high-throughput revision, with suggestions adopted instrumentally rather than evaluated for rhetorical fit and accuracy. Thus, chatbots may facilitate fluency and persistence while simultaneously increasing the risk of

shallow uptake unless learners are scaffolded to justify, reject, and strategically adapt system feedback.

Evidence from automated writing evaluation (AWE) research, to which chatbot-mediated feedback is increasingly closely aligned, reinforces this concern. Studies consistently show that AWE effects are heterogeneous and depend on the depth of learners' engagement with feedback rather than on tool presence alone (Chen & Cheng, 2008; Dikli & Bleyle, 2014; Grimes & Warschauer, 2010). Consequently, scholarship has examined learners' cognitive and affective processing during revision. Zhang and Hyland (2018) demonstrated that productive engagement requires emotional regulation alongside strategic processing, while Bai and Hu (2017) showed that uptake is selective and contingent on perceived accuracy. Ranalli (2018, 2021) further highlighted feedback clarity, proficiency, and trust as key moderators of uptake. Synthesised, this literature indicates that automated feedback supports development primarily when learners enact feedback literacy, evaluating suggestions, articulating revision rationales, and aligning changes with communicative intent.

Building on this body of work, the present study conceptualises chatbot-mediated feedback as a contingent pedagogical resource whose impact depends on learners' engagement with feedback across drafts and across feedback sources. By examining learners' uptake and revision reasoning over time, the study responds to calls for process-sensitive evidence on how AI-mediated feedback is interpreted, negotiated, and transformed into writing development rather than treated as a uniform instructional advantage.

### *Web-Based and AI-Supported Platforms in EFL Writing Development*

The diffusion of networked computing has reconfigured foreign language pedagogy, shifting writing instruction from static, product-oriented routines toward process-oriented activity mediated by digital platforms. Over the past few decades, the integration of computer-based technology into teaching and learning has expanded language education. Technology has become a routine component of learners' academic practice and a central concern for language instructors (Dash & Mohapatra, 2025). Numerous studies have shown that web-based platforms can support EFL/ESL writing development; however, effects are contingent rather than inherent. Specifically, platform-mediated gains are most consistently reported when tools reorganise the writing process, making revision trajectories visible, enabling iterative drafting, and scaffolding attention to higher-order concerns, rather than merely digitising conventional activities; this distinction is also emphasised in European open-learning discussions of cloud-based smart technologies as mechanisms for redesigning participation and feedback circulation, rather than as delivery infrastructure alone (Papadakis et al., 2023a, 2023b).

Research on online peer-review platforms indicates that structuring peer feedback within web-based environments can strengthen writing outcomes. Shang (2019) found that focused digital peer review, guided by criteria and supported by an online interface, is associated with more strategic revision, particularly for content and organisation. Zheng et al. (2015) similarly reported that a cloud-based classroom using Google Docs enabled multiple feedback cycles with peers and teachers, with learners improving not only grammatical accuracy but also coherence and paragraph structure through iterative in-platform revision. Mak and Coniam (2008) likewise showed that wiki-mediated writing in Hong Kong secondary classrooms enhanced motivation, organisation, and revision awareness by fostering collaboration and shared responsibility. Across these studies, a convergent explanation is that shared digital spaces increase the visibility and traceability of comments, revisions, and decision points, thereby supporting accountability and sustained refinement. Comparable affordance–constraint analyses have also been documented in North American settings where Google Docs is embedded in blended,

face-to-face writing instruction (Wilson, 2023). At the same time, traceability does not guarantee depth: without explicit guidance, learners may privilege readily "fixable" surface edits while under-attending to argument quality, rhetorical coherence, and audience alignment.

The affordances of blogging as a vehicle for extensive writing practice have also been documented. Sun (2010) introduced blogs as an informal, audience-aware medium and observed improvements in fluency, idea elaboration, and writing confidence. The public nature of blog posts can strengthen writer identity, and the asynchronous format may lower affective barriers to risk taking. Similar audience-mediated dynamics have been reported in tertiary EFL blogging, where perceived readership expands but participation can be uneven (Montalbán-Martínez, 2023). In contrast to tightly bounded classroom tasks, blogs can extend writing across time and open it to a wider audience, which may encourage sustained participation and richer idea development. However, that same public visibility can heighten evaluative pressure and widen participation gaps, so outcomes are likely shaped by learners' confidence, perceived scrutiny, and community norms.

Cloud-based collaborative writing further illustrates how real-time co-editing can reshape composing practices. Yim, Warschauer, Zheng, and Lawrence (2014) documented that students writing in Google Docs benefited from immediate peer/teacher feedback and engaged more rigorously with argument development and transition planning as they negotiated meaning and structure within a shared document. These findings suggest that peer-mediated, web-based platforms can promote collaboration, iterative drafting, and learner agency. However, agency should be treated as an instructional achievement rather than a platform default. Co-editing can distribute responsibility productively, but it can also reproduce unequal participation or superficial consensus unless roles, criteria, and reflective prompts are designed into the activity.

Building on these innovations, scholarship on technology integration emphasises that the pedagogical significance of tools lies in how they reorganise instructional activity. Technology integration has been defined in terms of how teachers use technology to perform familiar activities effectively and how technology reshapes these activities (Hennessy, 2005). Siyam et al. (2025) similarly define integration as the purposeful use of technology to improve the learning environment, enabling learners to complete tasks through digital platforms rather than exclusively through paper-based formats. Patel (2013) argues that technology can increase engagement and productivity, while multimedia resources can enrich linguistic knowledge by providing diverse materials for analysis and interpretation across contexts (Arifah, 2014). Taken together, this literature supports a distinction between technology as a delivery channel and technology as a mechanism for redesigning interaction, feedback circulation, and revision work, an analytic that is essential for interpreting the mixed findings in platform-based writing research.

### *Influence of Human and AI Feedback in EFL Writing Instruction*

Recent research has increasingly examined the comparative effects of human and AI-generated feedback in EFL writing instruction. Both feedback types afford distinct forms of support: human feedback is typically valued for contextual sensitivity, interpretive nuance, and affective scaffolding, whereas AI feedback is frequently characterised by immediacy, consistency, and availability. The incorporation of chatbot-mediated feedback into writing pedagogy has therefore introduced both opportunities and constraints in shaping learners' writing performance, engagement, and revision practices. Taken together, this literature frames human–AI comparisons not as a simple question of superiority but as a contingent relationship, in which outcomes vary with (a) the feedback's focus (surface accuracy versus meaning-level development), (b) learners' feedback literacy, and (c) the extent to which task prompts require learners

to justify accepting, rejecting, or adapting feedback, a framing that resonates with a recent study about comparisons of teacher and ChatGPT feedback quality on student essays, which foreground differences in prioritisation and criteria alignment (Steiss et al., 2024).

Human feedback has long been regarded as a central component of effective writing instruction. It commonly includes both surface-level corrections (e.g., grammar and lexis) and meaning-level commentary (e.g., content development, argumentation, and audience awareness) (Hyland & Hyland, 2006). Hyland and Hyland (2006) further emphasise the dialogic nature of teacher feedback, through which learners can seek clarification, negotiate meaning, and receive contingent scaffolding. Ferris (2011) similarly underscores the role of teacher feedback in supporting durable writing development, particularly when guidance is targeted, formative, and oriented toward learner autonomy. Nevertheless, human feedback should not be idealised as uniformly effective: its quality and focus can vary with workload, time constraints, and pedagogical priorities, resulting in differential depth, consistency, and timeliness. Accordingly, "human feedback" can be conceptualised as a pedagogical resource with well-established strengths, but constrained by limits on scalability and timeliness.

AI-generated feedback (e.g., ChatGPT) is typically rapid and standardised, often prioritising grammatical accuracy and sentence-level form over content-specific development (Chen et al., 2024; Bai & Hu, 2017). Although it can help learners identify linguistic errors, studies suggest it is less reliable for higher-order concerns such as organisation, coherence, and critical reasoning (Zou et al., 2025; Ranalli, 2021). Related evaluative work in higher-education contexts cautions that GPT-4-style feedback may be strongest for accuracy-oriented written corrective feedback unless prompts and criteria explicitly foreground discourse-level goals (Pack et al., 2025). For instance, Zou et al. (2025) found that students relied on teacher feedback for substantive revision while using ChatGPT mainly for surface editing, and Chen et al. (2024) reported frequent rejection of AI feedback perceived as vague, irrelevant, or excessive. Overall, AI feedback may increase revision speed and quantity but also encourage local "patching" unless learners are prompted to evaluate rhetorical fit and communicative intent, underscoring the value of examining revision processes and uptake rather than product outcomes alone.

Despite these limitations, AI feedback can enhance revision fluency and learner confidence, particularly when learners are trained to critically evaluate suggestions and justify revisions. Escalante et al. (2023) found no significant performance differences between human- and AI-reviewed groups, though students valued AI for speed while relying on teachers for depth and contextual specificity. This suggests that effectiveness depends less on feedback source than on learners' uptake over time, which is shaped by perceived clarity, tool credibility, and fit with proficiency and goals (Zhang & Hyland, 2018; Ranalli, 2018, 2021). Trust can facilitate uptake but is constrained by perceived accuracy and relevance, indicating that it must be calibrated to avoid uncritical adoption or premature dismissal; hence, instruction should cultivate principled criteria for evaluating both human and AI feedback and making defensible revision decisions; process-oriented evidence from L2 writing interventions integrating GenAI into revision work further underscores the importance of making deliberation and justification visible, rather than treating AI output as directly actionable (Michel et al., 2025).

Overall, recent scholarship suggests that human and AI feedback can be complementary when integrated strategically. Blended models that combine the interpretive depth of human commentary with the immediacy of AI support may be particularly well suited to heterogeneous EFL classrooms (Dizon et al., 2025; Klipp & Reinders, 2025). Building on this premise, the present study adopts a diachronic perspective to examine how learners' engagement with instructor feedback across drafts and to compare instructor feedback with post hoc GPT-4

feedback applied to the same baseline texts for analytic purposes. In doing so, the analysis traces observed revision patterns while using AI feedback as a comparative benchmark of feedback focus rather than as an input shaping students' revisions. The next section details the methodology, including participants, data collection, and analytic procedures.

## Methodology

### *Sampling*

This study involved twenty-six undergraduate EFL learners, all aged twenty or older, who completed a one-year foundational writing course primarily focused on grammar and paragraph-level writing. At the time of data collection, students had limited experience with academic essay writing, particularly in the five-paragraph format. Importantly, all student writing samples were collected prior to the public release of ChatGPT, ensuring that participants had no prior exposure to AI writing assistants. Students were asked to compose essays as part of regular classroom instruction and to record all revisions they made between drafts. These texts provide the basis for a diachronic analysis, in which writing development is examined across successive versions of the same task. Crucially, because the classroom writing cycle occurred before ChatGPT was publicly available, students' revisions reflect responses to instructor feedback only; AI-generated feedback was not part of the instructional process.

### *Research Context*

Although the student data were collected before ChatGPT became widely accessible, the present study retrospectively integrates AI-generated feedback using GPT-4 to examine its potential as a supplementary instructional tool. For analytic comparison, GPT-4 was applied to students' original drafts to generate feedback that could be directly compared with instructor comments. Because this AI feedback was generated post hoc and was not provided to students, it did not inform any revision decisions. Accordingly, the diachronic component of the study traces how students' texts changed across drafts in response to instructor feedback, while GPT-4 feedback serves as a post hoc benchmark for comparing feedback focus, coverage, and potential instructional utility. To avoid redundancy, a general description of GPT-4's capabilities is omitted; this section focuses on its study-specific use as a post hoc analytic comparator.

Ethical considerations for secondary analysis were addressed as follows. Students were informed that their course writing might be used for research purposes, and permission to use anonymised texts was obtained prior to analysis. Participation was voluntary, and students were provided an opt-out option without penalty. All essays were anonymised before analysis, and identifying information was removed from the dataset. Given the classroom-based and retrospective nature of the dataset, these procedures are reported to document responsible handling of student texts in the present analysis.

### *Types of Data*

The data for this study consist exclusively of student-produced written texts collected from February 2022 to June 2022, prior to the public release and classroom availability of ChatGPT. These texts are categorised as follows:

(a) Draft 1 (original drafts):
Students composed one essay genre: comparison. These drafts represent students' unassisted writing and serve as baseline texts for analysis.

(b) Draft 2 (drafts with feedback):
Each original draft was subjected to two forms of feedback: human-provided feedback from the instructor and AI-generated feedback using GPT-4. The AI feedback was produced post hoc and did not influence students' revisions; it is included solely to enable a controlled comparison with instructor feedback. Students' observed revisions were made only in response to instructor feedback during the course. Thus, the analysis contrasts (i) actual revision trajectories following human feedback with (ii) the content and emphases of GPT-4 feedback that could have been available under a counterfactual condition.

**Table 1.** *Types of data*

| Data collection method | Interpretation |
|---|---|
| 1. Students' original drafts | Essays composed independently by students prior to the availability of ChatGPT. |
| 2. Revised drafts (human feedback only) | Revisions made by students based on instructor feedback. |
| 3. GPT-4-generated feedback | AI-generated feedback applied retrospectively to original drafts to compare with human input. |

As shown in Table 1, the study analyses variation across multiple drafts to investigate how human feedback influenced student writing and how AI feedback might have supported revision if it had been available. More precisely, the diachronic analysis is grounded in students' observed draft-to-draft changes under instructor feedback. GPT-4 feedback is analysed in parallel as a post hoc comparator to identify convergences, divergences, and potential gaps in feedback coverage. The diachronic approach allows for the identification of developmental patterns across different stages of writing, providing insights into the comparative strengths and limitations of each feedback source. This design supports comparative interpretation of feedback characteristics, but it does not permit causal claims about the effect of AI feedback on student revision because learners did not receive AI feedback during drafting.

### *Texts*

The text corpus consists of approximately 50 to 60 essays written by 26 EFL students in an academic writing class. Each student composed one genre of the essay, one comparison, resulting in a multi-draft dataset. The texts are categorised into three stages:

(1) Draft 1 (original drafts) – independent student writing before receiving any feedback;
(2) Draft 2 (revised drafts) – texts revised using instructor feedback only;
(3) GPT-4 Feedback – original drafts annotated with GPT-4 feedback, applied post hoc for comparative purposes.

These documents form the basis for a diachronic analysis of revision behavior, error correction, content development, and feedback uptake. Specifically, revision behaviour and uptake are examined diachronically with respect to instructor feedback (from Draft 1 to Draft 2), while GPT-4 annotations on Draft 1 are analysed as a comparative reference for what AI feedback

would have foregrounded. The study explores whether and how the nature of human feedback differs from that of AI, and what implications this has for writing instruction in EFL contexts.

Figure 2 illustrates the interface through which GPT-4 feedback was generated and presented. This visualisation serves as a reference point for understanding how AI-mediated comments were produced and integrated into the corpus.

### *Procedures*

This study was conducted in three key stages, all designed to support a diachronic analysis of students' writing development under instructor feedback. The original data were collected before the public release of ChatGPT, see Figure 2. GPT-4 feedback was generated retrospectively to compare its input with that of human instructors. The overall procedure is explained as follows and illustrated in Figure 3 (Napkin, 2025).

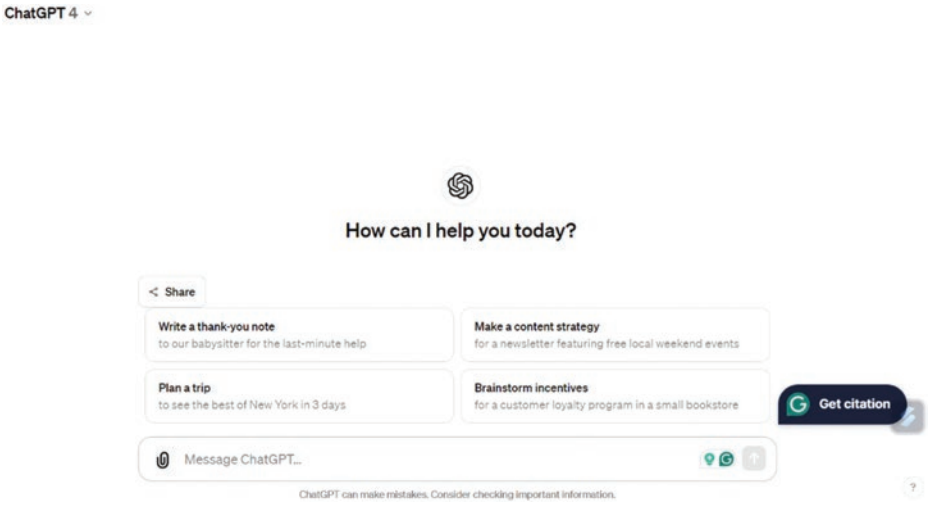


**Figure 2.** *Layout of GPT 4*

Figure 3 illustrates the sequential stages of the essay feedback and revision process implemented in the study. It begins with students' initial draft submission, followed by the provision of human feedback from instructors. Students then engage in revision based on feedback and subsequently submit a second draft. A peer comparison activity allows students to evaluate their writing in relation to their classmates' work. Subsequently, AI-generated feedback is applied to the original drafts, providing an additional layer of evaluation. Subsequently, after the classroom revision cycle was completed, the researcher generated GPT-4 feedback on the original drafts for comparative analysis; this AI feedback was not provided to students and did not inform any revision decisions. The process concludes with feedback comparison analysis, where researchers systematically examine and contrast the quality and instructional value of human and AI feedback.

#### Step One (Score One)

As part of their regular classroom activities, participants were instructed to compose a five-paragraph essay within the comparison genre. These essays, completed independently, served

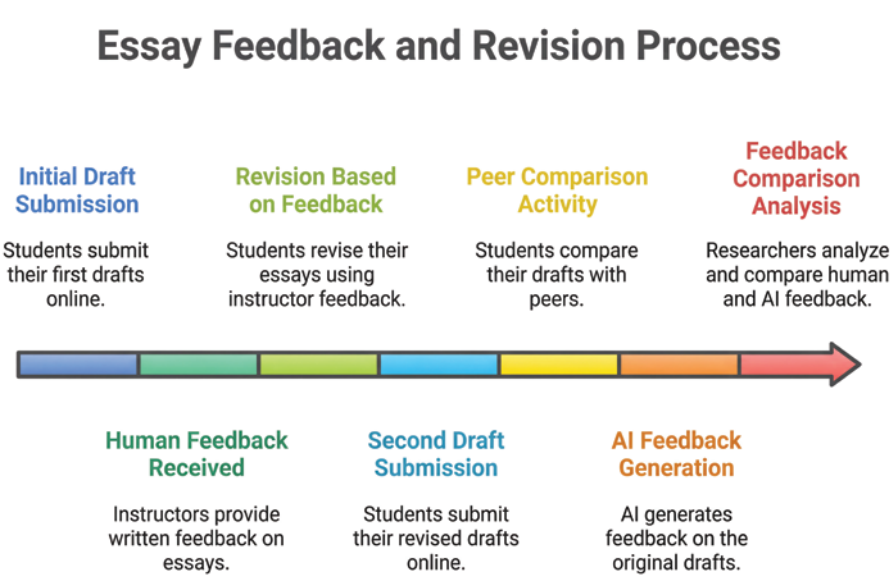


**Figure 3.** *Procedure of the study. Created using Napkin (Version 2025)*

as the students' original drafts. Participants were permitted to draft their essays at home, type them, and submit them via a web-based platform. Following the initial submission, students received human feedback and were required to revise their essays accordingly. The revised versions were then resubmitted and assigned a second score. Throughout this process, students were encouraged to consult and appropriately incorporate online sources to support their writing in a formal academic manner.

### Step Two (Score Two)

In the second stage, students revised their initial drafts based on feedback provided by the instructor and submitted a second version of their essays, which was then evaluated and assigned a new score. This revision process emphasised the incorporation of human feedback to improve content, structure, and clarity. Additionally, paired peer activities were implemented, allowing students to compare their original and revised drafts with those of a peer. These interactions fostered reflection on individual revision strategies and promoted greater awareness of how different feedback sources influenced their writing development.

### Step Three (Score Three)

In the final stage, conducted after ChatGPT became publicly available, the researcher employed GPT-4 to generate feedback on students' original drafts for comparative analysis. The AI-generated feedback addressed the same aspects as the instructor's comments, enabling a controlled comparison between the two feedback sources. Importantly, this stage was analytic rather than instructional: students did not receive, view, or respond to the AI-generated feedback at any point in the drafting process. Instead, the researcher independently analysed and compared the nature and focus of human and AI feedback. This analysis was then used to examine how students revised their drafts in response to instructor input, highlighting patterns of feedback uptake and revision behavior over time. Accordingly, the "diachronic" evidence in this study derives from the observed from Draft 1 to Draft 2 revision trajectory under instructor

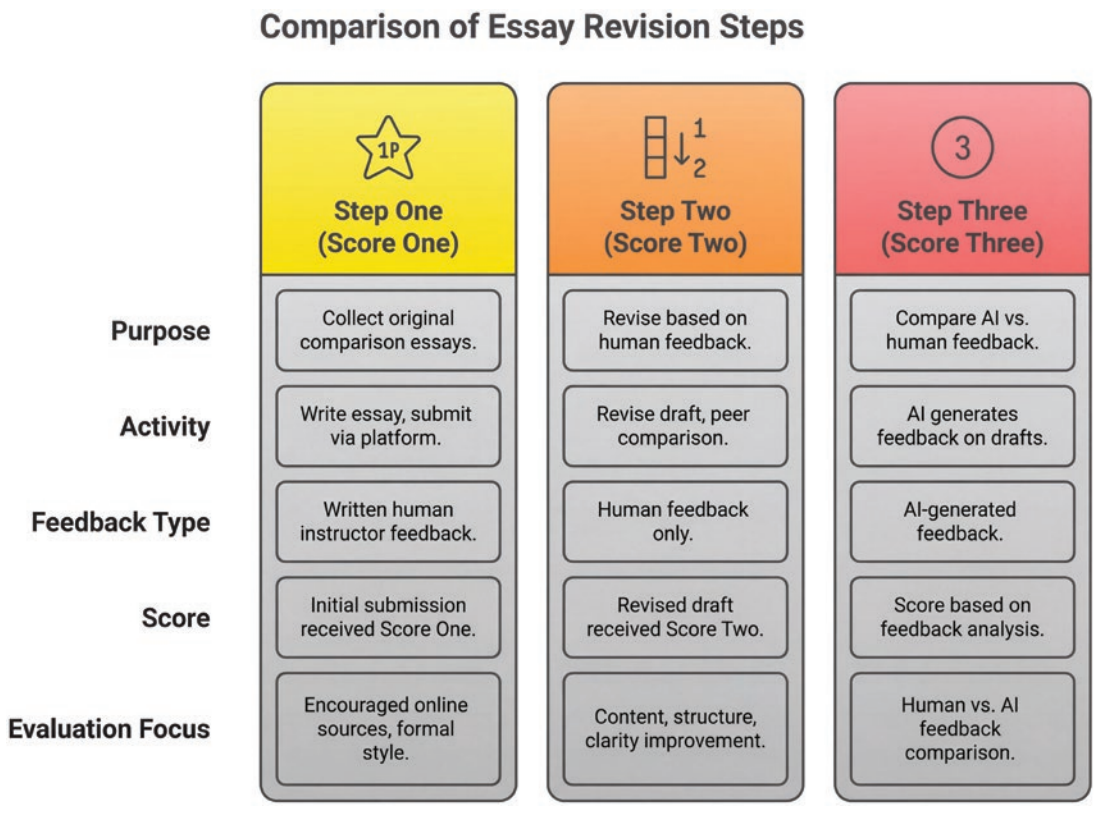


**Figure 4.** *Comparison of essay revision steps. Created using Napkin (Version 2025)*

feedback; GPT-4 feedback is used post hoc to support comparative interpretation of feedback focus and potential pedagogical affordances.

The analysis focused on student-produced written texts across multiple drafting stages, enabling a diachronic textual examination of feedback uptake, revision patterns, and linguistic development, as illustrated in Figure 4 (Napkin, 2025).

This diachronic design enables a controlled comparison of instructor and GPT-4 feedback anchored to the same baseline drafts, while tracing students' observed revisions under instructor feedback across Draft 1 and Draft 2. The next section details the analytical procedures used to examine the text corpus.

### *Analytical Approach*

The primary diachronic analysis examines how instructor feedback relates to students' observed revisions across drafts; AI-generated feedback (GPT-4) is analysed post hoc as a comparator rather than as a causal input to revision.

Data analysis focused on three sets of texts: (1) original drafts written independently, (2) revised drafts produced in response to instructor feedback, and (3) GPT-4-generated feedback applied retrospectively. By comparing these datasets, the researcher explored the types of changes students made, the depth and quality of revisions, and the kinds of feedback each source prioritised (e.g., surface-level vs. meaning-level features). Comparisons therefore address alignment and divergence between (a) the feedback characteristics of the two

sources and (b) the revision moves observed under human feedback, rather than attributing observed revisions to AI feedback.

The diachronic aspect of the study allowed the researcher to track developmental trajectories in students' writing under instructor feedback and to evaluate the extent to which human feedback and AI feedback differ in focus, specificity, and coverage when applied to the same baseline drafts. This analysis also included the classification of feedback comments (e.g., suggestion, correction, clarification) and the degree of alignment between the feedback type and the resulting revisions. Because GPT-4 feedback was generated after revisions were completed, "alignment" is interpreted descriptively (i.e., whether AI feedback targets issues that students later revised under human feedback), not as evidence of AI-driven uptake. The retrospective application of GPT-4 provided a valuable point of comparison for evaluating the potential instructional utility of generative AI in future classroom settings.

## Results

This study was situated in an intermediate, face-to-face academic writing course with 26 undergraduate EFL learners (n = 26). All student drafts were produced prior to the public release of ChatGPT; accordingly, Draft 1 and Draft 2 represent a human-feedback revision trajectory, whereas GPT-4 feedback was generated post hoc for comparative analysis.

The focal task in this phase was a comparison essay. Students selected one of three topics (Table 2), with most choosing "Life in colleges and in senior high school." After students submitted Draft 1, the instructor provided feedback and students revised to produce Draft 2; the same Draft 1 texts were subsequently evaluated by GPT-4 using an identical rubric (Ideas/Content 40%, Organisation/Structure 30%, Grammar/Language Use 30%). To address the reviewer's concern that score comparisons alone under-specify analytic claims, the Results section reports (a) score patterns (Table 4) and (b) a qualitative comparison of feedback focus using a brief coding scheme and illustrative excerpts drawn from GPT-4 feedback on the same Draft 1 texts.

**Table 2.** *Topics of the comparison essay*

| Possible topics | Number of students selected this topic | Students wrote the assigned topic |
|---|---|---|
| Life in colleges and in senior high school | 20 | 20 |
| Desktop computers and Tablet computers | 3 | 1 |
| Compare the cuisine of one country with the cuisine of another country | 3 | 3 |

Among the three possible topics, twenty students chose the compose the first topic 'life in colleges and in senior high school', while three students decided to write 'desktop computers and tablet computers' and three students composed 'compare the cuisine of one country with the cuisine of another country.' Students first received comments from the instructor. Subsequently, the same compositions were marked by AI. The prompt that the researcher applied to students' writings is run below (see Table 3).

**Table 3.** *GPT-4 prompts*

| |
|---|
| You have been assigned the task of evaluating an article that explores the comparison between life in college and life in senior high school. Your evaluation will be based on specific criteria outlined below. It is crucial to carefully read and understand these instructions. Please keep this document accessible during your review and refer to it as necessary.<br><br>*Evaluation Criteria:*<br>– Ideas and Content (40%):<br>Assess the strength and depth of the article's ideas. Consider the content provided and the overall coherence of the author's viewpoint. Evaluate the use of facts, data, or anecdotes to enhance the article's credibility.<br>– Organisation and Structure (30%):<br>Evaluate how well the article is organised. Check if the author contains four or five paragraphs (depending on the chosen method: point by point or block method) and if ideas flow logically. Assess the transitions between paragraphs and ensure that the article adheres to the 300-word limit or exceeds it.<br>– Grammar and Language Use (30%):<br>Examine the author's use of language. Evaluate if it is clear, precise, and appropriate. Determine if complex ideas are explained well and if the writing maintains reader interest. Pay attention to grammar, editing style, and ensure that only English words are used.<br><br>*Evaluation Steps:*<br>Read the following two example essays carefully.<br>Assess and summarise your evaluation for the student's essay only.<br>Provide an overall assessment of the article's effectiveness in presenting and supporting the statement.<br>Provide the final comprehensive score out of 100 in the form of "Final score: <score>".<br>Student's Essay:<br><br>[Essay text here] |

#### *Score Patterns across Draft 1, Draft 2, and GPT-4 Feedback*

Table 4 reports three values: Score 1 (Draft 1), Score 2 (Draft 2; instructor-guided revision), and GPT-4's post-hoc score on Draft 1. Overall, students' Draft 2 scores increased relative to Draft 1 scores, consistent with uptake of instructor feedback during revision. GPT-4 scores, generated on Draft 1 texts, are interpreted as a comparative benchmark of feedback orientation rather than as evidence of revision impact.

#### *Coding Categories for Feedback Type (Human vs. AI)*

Following prior feedback research, comments were coded along two coarse-grained dimensions that are directly interpretable for EFL writing pedagogy: (i) scope (local vs. global) and (ii) move type (directive vs. facilitative). To capture register-related issues observed in student drafts, an additional category was added for academic style/register (e.g., colloquial quantifiers).

- Local–Directive: points to an error and specifies a correction (e.g., grammar/wording).
- Local–Facilitative: flags a local issue but prompts the writer to revise (e.g., "rephrase for clarity").

**Table 4.** *Values: Draft 1, Draft 2 and GPT-4 post-hoc score*

| Student | Score 1 (Draft 1) | Score 2 (Draft 2) | Score evaluated by GPT4 | Remark |
|---|---|---|---|---|
| S1 | 78 | 83 | 81 | |
| S2 | 78 | 81 | 70 | |
| S3 | 49 | 71 | 70 | |
| S4 | 70 | 76 | 64 | |
| S5 | 76 | 82 | 78 | |
| S6 | 24 | 0 | 66 | |
| S7 | 68 | 80 | 74 | |
| S8 | 73 | 78 | 79 | |
| S9 | 55 | 70 | 70 | |
| S10 | 67 | 82 | 76 | |
| S11 | 52 | 76 | 72 | |
| S12 | 32 | 71 | 0 | Not the assigned topic |
| S13 | 70 | 78 | 82 | |
| S14 | 47 | 71 | 0 | Not the assigned topic |
| S15 | 74 | 78 | 79 | |
| S16 | 68 | 74 | 79 | |
| S17 | 66 | 80 | 79 | |
| S18 | 75 | 80 | 82 | |
| S19 | 74 | 78 | 82 | |
| S20 | 79 | 81 | 82 | |
| S21 | 70 | 76 | 82 | |
| S22 | 71 | 77 | 79 | |
| S23 | 67 | 77 | 77 | |
| S24 | 74 | 84 | 83 | |
| S25 | 76 | 80 | 86 | |
| S26 | 82 | 85 | 88 | |

- Global–Directive: specifies a discourse-level change (e.g., “add an example to support the claim”).
- Global–Facilitative: prompts reflection on argumentation/organisation (e.g., “consider strengthening the conclusion”).
- Register/Academic style: flags colloquial or vague expressions that weaken academic tone (e.g., repeated “a lot of”).

### *Qualitative Evidence: GPT-4 Feedback Excerpts and the Text Features*

Across essays, GPT-4 recurrently emphasised the need for evidence and specificity in the Ideas/Content domain. For example, GPT-4 noted that an essay "could benefit from more concrete examples, data, or anecdotes" to strengthen credibility, and similarly observed in another script that the argument "lacks specific examples, data, or anecdotes" that would enrich the comparison. These comments were coded as Global–Directive because they specify an instructional action (add evidence) linked to rubric criteria.

In Organisation/Structure, GPT-4 comments frequently targeted coherence at paragraph boundaries, often phrased as transition work. In one case, GPT-4 indicated that "ideas sometimes flow awkwardly … and the transitions could be smoother," while in another it recommended more "fluid" transitions and stronger synthesis in the conclusion. Such feedback was coded as Global–Facilitative when framed as a prompt to improve coherence rather than a specific rewrite.

For Grammar/Language Use, GPT-4 feedback tended to be local and pointing to concrete problematic strings. For instance, it flagged "minor grammatical errors and awkward phrases," exemplifying the issue with a quoted segment ("These fundamental knowledge is mostly learnt throughout high school"), and in other essays it similarly highlighted awkward phrasing and proofreading needs. These were coded as Local–Directive because they identify a linguistic problem and imply correction/editing.

### *Register Evidence: "a lot of" as a Recurrent Colloquial Quantifier*

In addition to grammar, several Draft 1 texts displayed register features that can depress perceived academic tone even when meaning is intelligible. One salient pattern was repeated reliance on the vague quantifier "a lot of", as illustrated in a student script: "spend a lot of time… covers a lot of content… spend a lot of time studying." In the present analysis, such instances were coded as Register/Academic style because the phrase is semantically underspecified and stylistically colloquial in academic prose; revisions typically require either quantification (e.g., "three hours daily") or more formal alternatives (e.g., "considerable," "substantial," "a significant amount of").

## Discussion

This section interprets the findings in relation to the study's process-oriented framing of feedback uptake and revision development. Consistent with the methodological note in Section 3, observed draft-to-draft changes are attributable to instructor feedback, whereas GPT-4 feedback functions as a post hoc comparison of feedback focus and potential pedagogical affordances.

### Response to RQ1

Instructor feedback and post hoc GPT-4 feedback overlapped most in identifying local language problems (e.g., grammar and sentence-level clarity). They diverged more on higher-order concerns: GPT-4 tended to provide rubric-driven, standardised advice (e.g., calls for examples and smoother transitions), whereas instructor feedback more consistently reflected context-sensitive judgement about task fulfilment and rhetorical appropriacy, especially when drafts deviated from the prompt.

### Response to RQ2

Students' Draft 1 and Draft 2 changes reflect uptake of instructor feedback, with the most consistent revisions occurring in areas where comments were specific and actionable (often local edits covering grammar, word choice or mechanics). Alignment with post hoc GPT-4 targets appears strongest for surface-level issues, while overlap is weaker for meaning-level concerns, suggesting that global revision work depends more heavily on teacher mediation and task-specific guidance.

Adopting a diachronic lens, this study examined EFL students' observed writing development from initial drafts (Draft 1) to instructor-supported revisions (Draft 2), and profiled how GPT-4 would have evaluated the same Draft 1 texts when applied retrospectively for comparative purposes. Consistent with prior work reporting broad convergence between human and AI evaluation on some outcome measures (Escalante et al., 2023), GPT-4's retrospective scores were generally comparable to instructor scores for many on-task drafts, particularly for surface-level accuracy. However, both the score patterns and the qualitative evidence reported in Section 4 indicate sharper divergence for higher-order dimensions (idea development and content relevance). For instance, when students deviated markedly from the prompt (e.g., S12, S14), GPT-4 issued a zero, reflecting rigid prompt adherence. In contrast, human instructors still provided constructive feedback, demonstrating greater flexibility and sensitivity to learner needs. This contrast aligns with research suggesting that learners often rely on human feedback for meaning-level issues while using AI support more selectively for local editing and, in some cases, organisational refinement (Jiang & Yu, 2024; Lin & Yang, 2011).

Diachronic evidence in this study derives from students' observed Draft 1 and Draft 2 revisions under instructor feedback, capturing how learners operationalised teacher guidance in their texts. GPT-4 feedback, applied post hoc to the same initial drafts, functions as a comparative reference for what the AI would likely foreground, typically local form and rubric-aligned clarity, but does not provide evidence of learner uptake. Accordingly, the diachronic comparison is not a test of AI "effectiveness." Rather, it clarifies which aspects of writing development appear to require contextualised mediation and how those needs might be anticipated within blended feedback designs.

Importantly, the qualitative excerpts in Section 4 provide a text-grounded account of these differences beyond score comparisons. Across scripts, GPT-4 frequently prioritised local form and rubric-aligned clarity, whereas instructor feedback more often incorporated context-sensitive guidance about task interpretation, elaboration, and rhetorical appropriacy. For example, GPT-4 repeatedly prompted writers to add "concrete examples, data, or anecdotes" and to strengthen coherence and transitions, while also highlighting grammar and awkward phrasing. Such feedback profiles are consistent with prior findings on AWE-style feedback being precise yet sometimes generic and decontextualised (Mekheimer, 2025). The register-related pattern observed in Section 4 (e.g., heavy reliance on colloquial quantifiers such as "a lot of") further illustrates how a feedback comparison can surface stylistic features relevant to academic writing quality that may not be fully captured by holistic scores alone.

Analytically, the contrast between instructor feedback and post hoc GPT-4 feedback is best interpreted through a feedback-literacy lens: what matters is not only comments but the extent to which feedback supports evaluation, reasoning, and goal-directed revision. In this dataset, GPT-4's rubric-driven comments functioned primarily as information, whereas instructor feedback more frequently functioned as mediation, narrowing attention to task-specific priorities and specifying what counts as an improved response in context. This distinction helps explain why overlap between sources was strongest for local accuracy but weaker for meaning-level development.

Overall, the evidence supports a complementary interpretation of human and AI feedback rather than a simple substitution model. AI feedback in this study was precise but often formulaic, tending to emphasise surface-level concerns such as grammar, spelling, and sentence clarity, a pattern mirrored in evaluations of automated writing evaluation (AWE) and Grammarly-like assistants (Mekheimer, 2025). Human feedback more often targeted context-specific elaboration and rhetorical guidance, particularly for content development. Accordingly, Draft 1 to Draft 2 improvements reflect instructor-guided revision, whereas the GPT-4 feedback is used post hoc to indicate what an AI system would likely prioritise if incorporated into instruction. This division of labour aligns with findings that learners show

stronger uptake and accuracy when revising with teacher feedback, while ChatGPT can be helpful for local editing and some organisational issues (Jiang & Yu, 2024; Escalante et al., 2023).

Pedagogically, the excerpt patterns in Section 4 suggest that blended feedback models are most sound when they (i) allocate AI tools to high-frequency, low-context tasks (e.g., recurring grammar, awkward phrasing, and register cues), and (ii) reserve instructor attention for meaning-level development (task fulfilment and rhetorical appropriacy). Importantly, the goal is not speed alone but engaged use. Students should be guided to explain why they accept, reject, or modify AI suggestions and how each change supports their intended meaning, so revision does not become many quick edits with limited improvement in overall quality.

Building on the comparative coding approach introduced in Section 4, the feedback profiles can be summarised along three rubric-aligned domains:

1. Ideas and Content: Human feedback frequently provided illustrative examples or suggestions for thematic elaboration, whereas AI tended to emphasise topic clarity and evidential support without offering extensive elaborative recommendations. This finding aligns with research noting that learners often perceive AI feedback on content as generic or misaligned with author intent (Chen et al., 2024).
2. Organisation and Structure: Both human and AI feedback similarly addressed issues related to paragraph coherence, structural organisation, and transitional phrases, displaying comparable efficacy partial convergence. In similar studies, students successfully integrated ChatGPT's organisational advice alongside teacher comments (Zou et al., 2025).
3. Grammar and Language Use: GPT-4 was notably more consistent and rigorous in identifying mechanical errors, while human instructors occasionally overlooked minor issues (Escalante et al., 2023; Mekheimer, 2025). In addition, the qualitative evidence suggests that some recurrent style issues (e.g., colloquial quantifiers) may be productively treated as a distinct "register" target in academic writing instruction.

An exploratory sub-analysis of sentence complexity was also conducted, grounded in Hunt's (1965) concept of T-units. While syntactic complexity features, such as subordinate clauses, nominalisation, and lexical density, are critical in academic writing, they were not explicitly integrated into the primary evaluation framework of this study.

An exploratory look at sentence complexity, using Hunt's (1965) T-unit heuristic, underscored that two sentences can be grammatically correct yet differ in sophistication (e.g., coordination vs. subordination). While syntactic complexity was not integrated into the primary scoring, Example 2, featuring subordination, demonstrates a higher level of integration, which AI may not consistently identify or address without targeted prompting. Related research suggests AI feedback can increase revision frequency and mechanical accuracy, but may require scaffolding to support deeper syntactic development and discourse-level choices (Escalante et al., 2023; Mekheimer, 2025). The following adapted examples (Table 5) illustrate differences in sentence complexity through the concept of T-units:

**Table 5.** *Analysis of complexity in sentences*

| T-unit | Examples |
|---|---|
| Example 1-Two T-units each with one clause | \| He is a student \| and he just moved here.\| |
| Example 2-One T-unit with two clauses | \| He is a student :: who just moved here. \| |

Although both examples above are grammatically accurate, the latter demonstrates greater sentence integration and sophistication, characteristics that AI-generated feedback might not consistently detect without targeted prompting. Despite GPT-4's effectiveness in identifying grammatical errors, the system lacks pedagogical adaptability in recognizing abstract reasoning, cultural references, or nuanced writer intentions. These findings underscore the continued pedagogical importance of human feedback in EFL writing instruction, especially during the early stages of learners' academic literacy development. At the same time, the present study's design and evidence base impose several methodological constraints that should be acknowledged explicitly.

Several limitations qualify the interpretation and generalisability of the findings. First, instructor scoring and feedback were produced by a single rater. Because no second rater independently scored a subset of texts, the consistency of human scoring could not be cross-checked, which limits claims about score comparability. Second, GPT-4 feedback was generated retrospectively and was not used in students' revision decisions. Accordingly, Draft 1 and Draft 2 changes cannot be attributed to AI feedback; GPT-4 is analysed only as a post hoc comparator of feedback focus and potential pedagogical affordances rather than as an intervention. Third, the study did not collect student perception data, such as usefulness, trust, cognitive load, or affective responses, restricting the analysis to textual outcomes and feedback characteristics rather than learners' experiential engagement with feedback. Fourth, applying GPT-4 post hoc may introduce systematic bias because the model's outputs are contingent on the specific prompt and the model version used at the time of analysis. Different prompts or models may yield different feedback profiles. Finally, the study did not include experimental comparison groups such as an AI-only feedback condition, a human-only condition with matched time-on-task, or a blended condition delivered instructionally, which precludes causal claims about the relative effectiveness of feedback sources. In addition, the dataset was drawn from a single course context and a single essay genre, which may limit the transferability of the findings to other proficiency levels, tasks, or instructional settings.

Ethical considerations are also noted. The essays were produced as part of routine course assessment, and students were informed that anonymised course writing might be used for research with an opt-out option; all texts were de-identified prior to analysis. Given the secondary-use nature of the dataset and the retrospective application of GPT-4, the post hoc comparisons should be interpreted cautiously as analytic profiling rather than evidence of instructional impact. Future research should incorporate multiple independent raters including a second instructor or trained coder to review scoring and feedback classifications and to resolve discrepancies through discussion. It should also include learner perception measures and controlled feedback conditions to test how AI feedback shapes revision when delivered in real time and supported by feedback-literacy instruction.

## Conclusion

Research in technology-enhanced language learning, particularly chatbot-supported writing instruction, has expanded rapidly, intensifying the need to examine how generative AI reshapes feedback practices and revision activity in EFL contexts. Positioned within this agenda, the present study adopted a learner-centred diachronic lens to examine students' observed Draft and Draft 2 revision trajectories under instructor feedback and to compare instructor feedback with GPT-4 feedback applied post hoc to the same baseline drafts. This design foregrounds not only writing outcomes but also the feedback features that are likely to matter for pedagogical integration.

In summary, the results indicate partial alignment between GPT-4 and instructor feedback, particularly in identifying recurrent local issues (e.g., grammar, wording, and sentence-level clarity) and in flagging general needs for coherence and transitions. However, the two sources diverged more clearly on higher-order dimensions: instructor feedback more consistently incorporated context-sensitive judgement about task fulfilment, rhetorical appropriacy, and elaboration, whereas GPT-4 feedback, while often detailed, tended to be more rubric-driven and could be less flexible when drafts deviated from prompt requirements. Importantly, students' observed improvements from Draft 1 to Draft 2 were shaped by instructor feedback because GPT-4 feedback was generated retrospectively and did not enter the classroom revision cycle. These findings support blended feedback models in which AI is positioned as supplementary support for rapid local editing and initial diagnostic input, while teachers retain responsibility for meaning-level guidance, task interpretation, and feedback mediation that develops learners' feedback literacy.

Rather than treating AI feedback as uniformly beneficial, the study highlights the conditions under which its strengths are most plausibly realised in instruction. AI tools can offer timely, standardised input that may reduce the burden of surface-level editing and support revision fluency; however, such support is pedagogically productive only when learners are guided to evaluate suggestions, justify revisions, and attend to communicative intent. The study also provides a foundation for further research on how AI-mediated feedback may be integrated into writing pedagogy without obscuring the role of instructional design. Future work should examine learners' cognitive and affective engagement with AI feedback through perception measures such as surveys, interviews, or stimulated recall and test blended feedback implementations in real-time classroom settings. Such work would clarify the conditions under which AI feedback promotes substantive revision rather than merely increasing surface-level editing.

In conclusion, this study contributes to AI-assisted language learning research by (i) documenting instructor-guided diachronic revision (Draft 1 to Draft 2) and (ii) offering a text-grounded, rubric-aligned comparison of instructor and GPT-4 feedback characteristics. The findings underscore that AI is best framed as a complementary tool rather than a replacement for human instruction, particularly in tertiary EFL writing contexts where meaning-level guidance and contextual judgement remain essential. Pedagogically, teachers and curriculum designers are encouraged to integrate AI tools strategically—using AI to support local editing and feedback access while maintaining human oversight for rhetorical development and for cultivating critical digital literacy in responding to automated suggestions.

## References

Ahmadi, R. M. (2018). The use of technology in English language learning: a literature review. *International Journal of Research in English Education*, *3*, 115–125. https://doi.org/10.29252/ijree.3.2.115

Alsaleem, B. I. A. (2014). The effect of "WhatsApp" electronic dialogue journaling on improving writing vocabulary, word choice, and voice of EFL undergraduate Saudi students. In 21st Century Academic Forum Conference Proceedings. Harvard. https://www.researchgate.net/publication/349830602

Arifah, A. (2014). *Study on the use of technology in ELT classroom: Teachers' perspective* (Master's thesis, BRAC University, Dhaka, Bangladesh). BRAC University Institutional Repository. https://dspace.bracu.ac.bd:8443/xmlui/bitstream/handle/10361/3999/12363008.pdf

Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: How do students respond? *Educational Psychology*, *37*(1), 67–81. https://doi.org/10.1080/01443410.2016.1223275

Belda-Medina, J., & Calvo-Ferrer, J. R. (2022). *Using chatbots as AI conversational partners in language learning*. *Applied Sciences*, *12*(17), 8427. https://doi.org/10.3390/app12178427

Budiu, R. (2018, November 25). *The user experience of chatbots*. Nielsen Norman Group. https://www.nngroup.com/articles/chatbots/

Carayannopoulos, S. (2018). Using chatbots to aid transition. *The International Journal of Information and Learning Technology*, *35*(2), 118–129. https://doi.org/10.1108/IJILT-10-2017-0097

Chatbot. (n.d.). In *Wikipedia*. Retrieved August 16, 2025, from https://en.wikipedia.org/wiki/Chatbot

Chen, C.-F. E., & Cheng, W.-Y. E. (2008). *Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. Language Learning & Technology, 12*(2), 94–112. https://www.lltjournal.org/item/10125-44145/

Chen Z., Zhu X., Lu Q. & Wei W. (2024). L2 students' barriers in engaging with form and content-focused AI-generated feedback in revising their compositions, Computer Assisted Language Learning, https://doi.org/10.1080/09588221.2024.2422478

Dash, N., & Mohapatra, L. M. (2025). Technology empowered teaching and learning in language education: A systematic review. In *Technology driven language learning: Innovations and applications* (pp. 127–145). Springer. https://link.springer.com/chapter/10.1007/978-3-031-77232-0_8

Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing*, *22*, 1–17. https://doi.org/10.1016/j.asw.2014.03.006

Dizon, G., Gold, J., & Barnes, R. (2025). ChatGPT for self-regulated language learning: University English as a foreign language – students' practices and perceptions. *Digital Applied Linguistics, 3*, 102510. https://doi.org/10.29140/dal.v3.102510

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*(1). https://doi.org/10.1186/s41239-023-00425-2

Ferris, D. R. (2011). Treatment of error in second language student writing (2nd ed.). University of Michigan Press. https://doi.org/10.3998/mpub.2173290

Fryer, L., & Carpenter, R. (2006). *Bots as language learning tools*. *Language Learning & Technology, 10*(3), 8–14. http://dx.doi.org/10125/44068

Fryer, L. K., Nakao, K., & Thompson, A. (2019). Chatbot learning partners: Connecting learning experiences, interest and competence. Computers in Human Behavior, 93, 279–289. https://doi.org/10.1016/j.chb.2018.12.023

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. Journal of Technology, Learning, and Assessment, 8(6), 4–43. https://files.eric.ed.gov/fulltext/EJ882522.pdf

Heller, B., Proctor, M., Mah, D., Jewell, L., & Cheung, B. (2005). *Freudbot: An investigation of chatbot technology in distance education*. In *EdMedia+ Innovate Learning* (pp. 3913–3918). Association for the Advancement of Computing in Education (AACE). https://psych.athabascau.ca/html/chatterbot/ChatAgent-content/EdMediaFreudbotFinal.pdf

Hennessy, S. (2005). Emerging teacher strategies for supporting. University of Cambridge.

Hunt, K.W. (1965). Grammatical structures written at three grade levels (No.3). National Council of Teachers of English.

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. Language Teaching, 39(2), 83–101. https://doi.org/10.1017/S0261444806003399

Jia, J., Chen, Y., Ding, Z., & Ruan, M. (2012). Effects of a vocabulary acquisition and assessment system on students' performance in a blended learning class for English subject. Computers & Education, 58(1), 63–76. https://doi.org/10.1016/j.compedu.2011.08.002

Jia, J., & Ruan, M. (2008). Use chatbot CSIEC to facilitate the individual learning in English instruction: A case study. In B. P. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie (Eds.), Intelligent Tutoring Systems (pp. 706–708). Springer. https://doi.org/10.1007/978-3-540-69132-7_84

Kerly, A., Hall, P., & Bull, S. (2007). Bringing chatbots into education: Towards natural language negotiation of open learner models. Knowledge-Based Systems, 20(2), 177–185. https://doi.org/10.1016/j.knosys.2006.11.014

Klipp, J., & Reinders, H. (2025). Effects of different types of computer-assisted corrective feedback on L2 pragmatics learning. *Digital Applied Linguistics, 3*, 102515. https://doi.org/10.29140/dal.v3.102515

Kroll, B. (2001). Considerations for teaching writing in English as a second or foreign language. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (3rd ed., pp. 233–248). Boston, MA: Heinle & Heinle.

Lampropoulos, G., Papadakis, S. (2025). The educational value of artificial intelligence and social robots. In: Lampropoulos, G., Papadakis, S. (eds) Social robots in education. Studies in computational intelligence, vol 1194. Springer, Cham. https://doi.org/10.1007/978-3-031-82915-4_1

Lin, W., & Yang, S. (2011). Exploring students' perceptions of integrating Wiki technology and peer feedback into English writing courses. English Teaching: Practice and Critique, 10(2), 88–103. https://eric.ed.gov/?id=EJ944900

Mak, B., & Coniam, D. (2008). Using wikis to enhance and develop writing skills among secondary school students in Hong Kong. System, 36(3), 437–455. https://doi.org/10.1016/j.system.2008.02.004

Mekheimer, M. (2025). Generative AI-assisted feedback and EFL writing: a study on proficiency, revision frequency and writing quality. Discover Education, 4, 170. https://doi.org/10.1007/s44217-025-00602-7

Michel, M., Bazhutkina, I., Abel, N., & Strobl, C. (2025). Collaborative writing based on generative AI models: Revision and deliberation processes in German as a foreign language. *Journal of Second Language Writing*, *67*, 101185. https://doi.org/10.1016/j.jslw.2025.101185

Mntalbán-Martínez, N., García Pinar, A., & Tello Fons, I. (2023). Blogging in EFL: A writing project using ICT. *The EuroCALL Review*, *30*(2), 37–54. https://doi.org/10.4995/eurocall.2023.15722

Napkin. (2025). Napkin [Generative AI model]. https://napkin.ai

OpenAI. (2025). ChatGPT (Version 4) [Large language model]. https://chat.openai.com

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

Papadakis, S., Kiv, A. E., Kravtsov, H. M., Osadchyi, V. V., Marienko, M. V., Pinchuk, O. P., ... Semerikov, S. O. (2023). Revolutionizing education: Using computer simulation and cloud-based smart technology to facilitate successful open learning. In *Joint Proceedings of CoSinEi & CSTOE 2022* (CEUR Workshop Proceedings, Vol. 3358, pp. 1–18). https://doi.org/10.31812/123456789/7375

Papadakis, S., Kiv, A. E., Kravtsov, H. M., Osadchyi, V. V., Marienko, M. V., Pinchuk, O. P., ... Semerikov, S. O. (2023). Unlocking the power of synergy: The joint force of cloud technologies and augmented reality in education. In *Joint Proceedings of CTE 2021 & AREdu 2022* (CEUR Workshop Proceedings, Vol. 3364, pp. 1–23). https://doi.org/10.31812/123456789/7399

Patel, C. (2013). Use of multimedia technology in teaching and learning communication skill: An analysis. *International Journal of Advancements in Research & Technology*, *2*(7), 116–123.

Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning*, *31*(7), 653–674. https://doi.org/10.1080/09588221.2018.1428994

Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing,* 52. https://doi.org/10.1016/j.jslw.2021.100816

Schmulian, A., & Coetzee, S. A. (2019). The development of messenger bots for teaching and learning and accounting students' experience of the use thereof. *British Journal of Educational Technology, 50*(5), 2751–2777. https://doi.org/10.1111/bjet.12723

Shang, H. F. (2019). Exploring online peer feedback and automated corrective feedback on EFL writing performance. *Interactive Learning Environments*, *30*(1), 4–16. https://doi.org/10.1080/10494820.2019.1629601

Siyam, Y., Siyam, N., Hussain, M., & Alqaryouti, O. (2025). Evaluating technology integration in education: a framework for professional development. *Discover Education, 4*(53). https://doi.org/10.1007/s44217-025-00448-z

Smutny, P., & Schreiberova, P. (2020). Chatbots for learning: A review of educational chatbots for the Facebook messenger. *Computers & Education*, *151*, 1–10. https://doi.org/10.1016/j.compedu.2020.103862

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

Sun, Y. (2010). Extensive writing in foreign-language classrooms: a blogging approach. *Innovations in Education and Teaching International*, *47*(3), 327–339. https://doi.org/10.1080/14703297.2010.498184

Warschauer, M. (2000). The death of cyberspace and the rebirth of CALL. *English Teachers' Journal*, *53*, 61–67. http://www.gse.uci.edu/markw/cyberspace.html

Wilson, D. (2023). A collaborative story writing project using Google Docs and face-to-face collaboration. *Canadian Journal of Learning and Technology*, *49*(3). https://doi.org/10.21432/cjlt28174

Yim, S., Warschauer, M., Zheng, B., & Lawrence, J. F. (2014). Cloud-based collaborative writing and the common core. Journal of Adolescent & Adult Literacy, 58(3), 243–254. https://doi.org/10.1002/jaal.345

Zheng, B., Lawrence, J., Warschauer, M., & Lin, C.-H. (2015). Middle School Students' Writing and Feedback in a Cloud-Based Classroom Environment. *Technology, Knowledge and Learning*, *20*, 201–229. https://doi.org/10.1007/s10758-014-9239-z

Zhang, Z., & Hyland, K. (2022). Fostering student engagement with feedback: An integrated approach. *Assessing Writing,* 51. https://doi.org/10.1016/j.asw.2021.100586

Zorko, V. (2007). A rationale for introducing a wiki and a blog in a blended-learning context. *CALL-EJ online*, *8*(2). Retrieved from http://callej.org/journal/8-2/zorko.html

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*, 1–30. https://doi.org/10.1080/09588221.2024.2447279