



Castledown

 OPEN ACCESS

Technology in Language Teaching & Learning

ISSN 2652-1687

<https://www.castledown.com/journals/tltl/>

Technology in Language Teaching & Learning, 7(1), 2208 (2025)

<https://doi.org/10.29140/tltl.v7n1.2208>

Exploring AI to Automate EFL Corrective Written Feedback in the First Language



ROB HIRSCHEL^a 

KAYOKO HORAI^b 

Sojo University, Japan
hirschel@m.sojo-u.ac.jp

Sojo University, Japan
khorai@m.sojo-u.ac.jp

Abstract

With the advent of generative AI that uses large language models such as ChatGPT, it is now relatively easy to provide automated written corrective feedback in a student's native language. This paper reports on an exploratory study using a ChatGPT-powered plugin recently developed for the Moodle learning management system. The classroom intervention on which this study is based had four main steps: a) brainstorming and vocabulary search (5 minutes), b) subsequent free-writing activity in an online textbox (10 minutes), c) reading the ChatGPT feedback in L1 Japanese (5 minutes), and d) completing paper error logs to process feedback (5 minutes). Data were collected from students' written submissions as well as the ChatGPT-generated feedback, students' error logs, and surveys administered to students after each activity and at the end of the semester. Analysis of survey data and ChatGPT-generated feedback suggested that both students and the teacher appreciated the grammatical feedback, found it to be relatively accurate, and largely comprehensible. There is room, however, for improvements in accuracy and presentation. Pedagogical implications are discussed.

Keywords: corrective feedback, AI, automated writing feedback, L2 writing

Introduction

Central to second language acquisition (SLA) are the constructs of input, output, feedback and noticing (Mackey, 2006; Long, 1996; Schmidt, 1990; Swain, 1995, 2000; Swain & Suzuki, 2008). There is general consensus that these constructs are the building blocks of language acquisition, though scholars

Copyright: © 2025 Rob Hirschel & Kayoko Horai. This is an open access article distributed under the terms of the Creative Commons Attribution Non-Commercial 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. **Data Availability Statement:** All relevant data are within this paper.

are not always in agreement as to whether all are fundamentally necessary, the timing of introduction, or the proportion of each. What is undisputed is that having access to the target language is essential.

In the past 30 years, the growth of the internet has exponentially increased the amount of comprehensible second language (L2) input available to students. One need only consider web pages, videos, and social networking sites among the many new affordances that now exist. The opportunity to produce comprehensible output has also been greatly facilitated. Written and video posts to social networking sites as well as online communication mean that language students of today can easily produce both written and spoken output, even in the physical absence of an interlocutor, a partial solution to the problem Kato (2020) defined as “poverty of opportunity.” Alongside this growth in input and output, there may also be a chance for increased feedback and noticing, perhaps in clarification requests from an online interlocutor. Nevertheless, actionable corrective feedback (CF) is something that is relatively rare outside the traditional classroom practices of recasts and clarification requests (for spoken production) and evaluated texts (for written production).

Even within the L2 classroom, however, CF can be sparse. Many EFL and ESL teachers do not provide extensive CF for reasons that include, but are not limited to, the considerable time required to provide it, the unhelpful but necessary delay in returning papers marked with CF, concerns about the comprehensibility of the feedback in the L2 and concerns that students may simply disregard feedback (Hirschel, 2011).

In recent years automated writing feedback (AWF), providing instantaneous corrective written feedback, has become widely available. In a meta-analysis of 83 studies focused on AWF through the year 2022, Shi and Aryadoust (2024) found that although a clear majority of studies were focused on writing and/or language study at the collegiate level, the AWF systems examined and the focus of those examinations were varied. The most common focus of the studies was on form-related feedback and the impact on writing performance, followed by examination of student and teacher perspectives, and engagement with the systems. Though the authors cautioned that AWF systems appeared incapable of sufficiently representing the complex construct of writing, the authors did express *excitement* at the opportunities that may be offered by generative AI such as ChatGPT. Shi and Aryadoust (2024) nevertheless warned that, for educators to appropriately select the AWF systems that best match their learners’ needs, AWF systems must be transparent about the training data used in their models.

In another meta-analysis of 82 studies on AWF (referred to as automated corrective feedback or ACF in their paper), Shadiev and Feng (2024) examined, among other aspects, the functions of 43 AWF tools, participant perceptions, and reported accuracy. The AWF tools varied in terms of the feedback focus (general writing, grammar, spelling, collocation and word use), in terms of feedback provided (direct, indirect, direct or indirect combined with teacher or peer feedback), in the learner’s ability to personalize settings (e.g., target specific errors) and in the formatting and display of feedback. In the 11 studies reporting some element of accuracy, grammatical accuracy varied widely (58% to 97.5%) with the large range impacted by the quality of the AWF tool, the type and level of learner writing under analysis, and the units of analysis. In the 35 studies that reported participants’ perceptions of the AWF program, 27 (77%) reported positive reactions. Advantages of AWF cited by Shadiev and Feng included time saved by teachers, instantaneous feedback, better understanding, increased motivation, and facilitating learner autonomy. Disadvantages included inaccuracies, overly vague feedback, and systems that were powered by insufficient data to consistently render effective feedback.

For some time now, commercial programs such as Grammarly (Windah et al., 2024), Criterion (Wolf & Oh, 2024), and ProWritingAid (Giessler, 2023) have provided AWF that is used in academic and non-academic settings alike. These programs potentially solve issues of teachers spending long hours

marking papers and the delay between students writing and receiving feedback. However, programs such as the aforementioned have primarily been designed for first language (L1) users, are primed to give feedback on common L1 errors, and provide feedback that is likely beyond the understanding of many English language learners (Godwin-Jones, 2022).

Literature Review

Generative AI has been described as a *paradigm shift* and something that is *here to stay* (Stockwell, 2024). Despite being warned that teachers ignore these technologies at their peril, Stockwell (2024) also noted that “at present, the evidence for learner development as a result of AI is tenuous” (p. 5). Another recent paper found that, even though independent English learners’ positive orientation to and frequent use of ChatGPT signifies AI as a powerful tool for the future, further empirical research is required (Liu and Ma, 2024). The papers in the next section demonstrate how developments in generative AI are increasingly enabling educators to realize the benefits of seemingly effortless, instantaneous and personalized feedback of the commercial programs referenced above with feedback that is better geared towards L2 learners.

Recent Studies Using Generative AI for Corrective Feedback

A recent study of upper-intermediate university EFL learners compared their perceptions of CF from three different sources: peers, ChatGPT, and the classroom teacher (Zeevy-Solovey, 2024). Though the students indicated that they completely understood the comments from all sources, feedback from the teacher ranked most popular, followed by feedback from both the teacher and ChatGPT. Students indicated that the teacher was the expert and, as such, the teacher’s feedback could be trusted and was useful. Both the researcher and some students suggested that ChatGPT could be helpful due to the speed of feedback, but concerns regarding consistency were expressed. The study was supportive of ChatGPT in its ability to further student autonomy, but cautioned that students need to be prepared to use the chatbot critically and “to formulate clear and detailed prompts” (p. 15).

In a late 2023 study of intermediate English learners at a Turkish university, student participants were asked to use ChatGPT for feedback on aspects such as register and grammar, with the researchers finding that, despite some technical difficulties and concerns about the chatbot’s ability to differentiate formal and informal registers, students were generally enthusiastic about the task and benefitted from the use of ChatGPT as a learning assistant (Özçelik & Ekşi, 2024). This study was limited, however, to a sample size of 11 participants all around the CEFRA B1 level. The researchers found that accurate prompt engineering was “crucial” (p. 14), and, in a point to be returned to later, student participants were instructed to use ChatGPT prompts provided by their teacher.

In a separate qualitative study at a different Turkish university, researchers determined that the 13 preparatory class students found ChatGPT to be useful for improving writing, vocabulary and grammatical knowledge, and efficient for research, but that students were also concerned by connectivity issues as well as the possibilities of overreliance on the tool and skills deterioration (Karataş et al., 2024). In this study, the student participants were initially provided by the teacher with ChatGPT prompts and were thereafter encouraged to develop their own prompts “to foster independence and creativity” (Section 2.5 Language learning activities).

Moving to the context of an American graduate school, researchers investigated the behavioural, cognitive, and affective engagement of six Iranian graduate students with ChatGPT feedback on their research proposals. Koltovskaia et al. (2024) found that the graduate students largely used just one of ten instructor-provided prompts, the one soliciting identification and correction of errors in grammar

and punctuation. The student participants appeared to cognitively engage with the feedback, accepting most of the accurate ChatGPT feedback and correctly rejecting the majority of inaccurate feedback. Affectively, students demonstrated positivity toward and appreciation for the tool, particularly in addressing matters of word choice and paraphrasing. In regard to potential concerns, student participants were occasionally frustrated that the feedback might not be completely reliable and one participant indicated apprehension as to the privacy of data entered into the platform.

Researchers Jiang et al. (2024) conducted a study focused mainly on critical thinking with both L1 and L2 learners of Chinese at a Canadian university. The researchers found that students successfully used ChatGPT for feedback on grammatical errors, questions of word choice, eliciting background information, questioning reasoning, and on other aspects of tasks that differed by class. In terms of facilitating critical thinking, Jiang et al. (2024) found that a higher level of Chinese proficiency did not correlate to higher order critical thinking. The researchers instead found that the type of task assigned and the type of prompts provided to students were more predictive of the level of critical thinking displayed by learners. In terms of pedagogical implications, it was determined that the *non-authoritative nature* of ChatGPT enabled students to critically assess feedback with scepticism, enabling the learners to develop better critical thinking skills. Educators were encouraged to more carefully design their classroom tasks in order to maximize the benefits of ChatGPT and minimize inaccuracies and potentially overlooked learning opportunities.

Another study of native Chinese speakers analysed use of a chatbot to provide corrective feedback on conversational English in two forms (guided self-correction versus explicit feedback) in two modalities: within the chatbot and in a separate GUI. Instead of using a commercially available generative AI program such as ChatGPT or Ollama, the researchers designed a custom-built chatbot that gave feedback on targeted errors for a textual conversation (typed interaction with the chatbot) on the predetermined topic of travel. Findings suggested that guided self-correction may provide for a better learning experience, particularly for lower-level learners and particularly for those with high learning motivation. In analysis of survey data, *serious learners* in the guided self-correction group appeared to experience higher levels of confidence with their grammar skills during the activity and also appeared to perceive the feedback as more useful. This exploratory study was limited, however, to five targeted error types with a prescribed conversation topic.

Concerns

Despite receiving individualized feedback, however, it appears that the student participants in at least four of the six above-mentioned studies received the feedback in the L2, an issue that may preclude many language learners from effectively engaging with the feedback. Certainly, AI is capable of rendering L2 feedback into the learner's L1, but there are still concerns that feedback, especially CF with grammatical metalanguage, could get *lost in translation* (Moneus & Sahari, 2024). A second concern is that successful generative AI prompt engineering requires a special skill set that, to achieve effective results, necessitates time, effort and practice (Jiang et al., 2024; Zeevy-Solovey, 2024). Until recently (see Green, 2024), use of generative AI has also required students to navigate outside their learning management system (LMS) or other online platform, a relatively easy step but nevertheless one more possible complication for students on top of the inevitable technical difficulties (Özçelik & Ekşi, 2024). Though it may be appropriate in some contexts for students to engage chatbots directly, students in typical required university EFL courses, as in the current study, are more likely to benefit from AI-generated feedback resulting from set prompts and models that have been carefully designed, tested, and finely tuned.

Concerns are not limited to the language of the CF, prompt engineering, or successful navigation through a new website. Security of data entered by learners is yet another issue, as raised by Koltovskaia et al.

(2024). Finally, there is the concern that Spring (2024) has called the “black box” issue with regard to platforms such as ChatGPT which are not open-source, where users do not know how the models respond to prompts, and where users have no knowledge upon what data the models are trained.

AI Text Plugin

A potential answer to at least some of these concerns is the recent development of a plugin that enables learners to benefit from chatbot-generated feedback without leaving the Moodle LMS or entering a prompt (Green, 2024). All students must do is type their text and click a ‘*check*’ icon. Within about 20 seconds students see feedback in their L1 provided by a generative AI platform that has been prompted, behind the scenes, by a teacher. At the time of writing, this plugin, entitled AI Text, supported ChatGPT and Ollama. The authors were drawn to the development of the AI Text tool as a way of potentially mitigating concerns about the language of the CF in the L2 (Karataş et al., 2024; Koltovskaia et al., 2024; Özçelik & Ekşi, 2024; Zeevy-Solovey, 2024), technical difficulties in navigating a new platform (Özçelik & Ekşi, 2024), concerns about prompt engineering or pasting a teacher-provided prompt (Jiang et al., 2024; Karataş et al., 2024; Koltovskaia et al., 2024; Özçelik & Ekşi, 2024; Zeevy-Solovey, 2024), as well as worries about privacy (Koltovskaia et al., 2024).

Method

The current study was grounded in an Exploratory Practice approach (Allwright, 2005) that sought to ethically understand a component of the language classroom, give a voice to both students and their classroom teacher, but encumber neither students nor the teacher with unsustainable demands. The component of classroom practice under consideration was a free writing activity that had been added to the curriculum a year earlier. The intervention incorporated ChatGPT-generated corrective feedback delivered in the students’ L1 (Japanese) such that students could effectively comprehend the nature of their errors and the proposed corrections. The specific research questions for this exploratory project, related to the new Moodle *AI Text* item-type, powered in this instance by ChatGPT (GPT-4/GPT-4o), were:

RQ1: Do students find the feedback provided by the AI Text item meaningful?

RQ2: Does the classroom teacher find the feedback provided by the AI Text item meaningful, particularly in regard to accuracy and clarity?

Participants

The participants comprised 84 second-year university students taking a compulsory English course in one of three Faculties: Art (28), Biotechnology and Life Sciences (BioLife, 27), and Engineering (29). The average student was generally at about the A1 level of the Common European Framework of Reference for Languages (CEFRA), as observed by the classroom teacher and suggested by scores on the institutional placement test. The participants’ first two semesters (year one) of required English education at the university focused almost entirely on oral communication with only incidental exposure to writing through tasks such as completing worksheets and typing up transcripts of audio-recorded conversations (Kirchmeyer, 2023). As far as we are aware, the students had minimal, if any, exposure to AWF tools. University-wide survey results suggested that while about 60% of students had experience using ChatGPT in any capacity, only 13% reported having used it for English study (Horai, 2025), many of whom were likely to have been the same participants using an API for ChatGPT in the current study. The first author, a native English speaker, was the classroom teacher for the three classes, while the second author, a Japanese teacher of English, assisted with AI prompt refinement and analysis of AI-generated feedback. The English classes met twice a week for 15 weeks during the

Spring 2024 academic semester. All participants gave informed consent both to submit their data to an OpenAI API and to have their data analysed for research purposes. The participants were advised that any use of their data would be anonymized and that their informed consent could be withdrawn at any time.

Procedure

The procedures were not as straightforward as had been anticipated. Originally intended as a simple qualitative analysis of student feedback to a classroom intervention, the authors found that steps for AI prompt refinement had to be added. There were, therefore, alternating steps between prompt engineering, classroom intervention, prompt revision and other necessary measures.

Classroom Intervention

The classroom intervention (see Figure 1) took place during the latter half of the semester and was conducted once a week over four or five weeks (depending on class) during the first 20–25 minutes of regular face-to-face classes. The intervention can most simply be described as an online version of a ten-minute free-writing activity that students had hitherto been doing on paper approximately once a week. During the first half of the semester, when students were completing the activity on paper, the teacher had later collected the papers, read them and provided short comments in green pen such as *Me Too!*, *Wow!* and ☺ (smiley faces) in response to written content. No CF was provided by the teacher. No instruction in writing conventions was given in these otherwise largely oral communication classes.

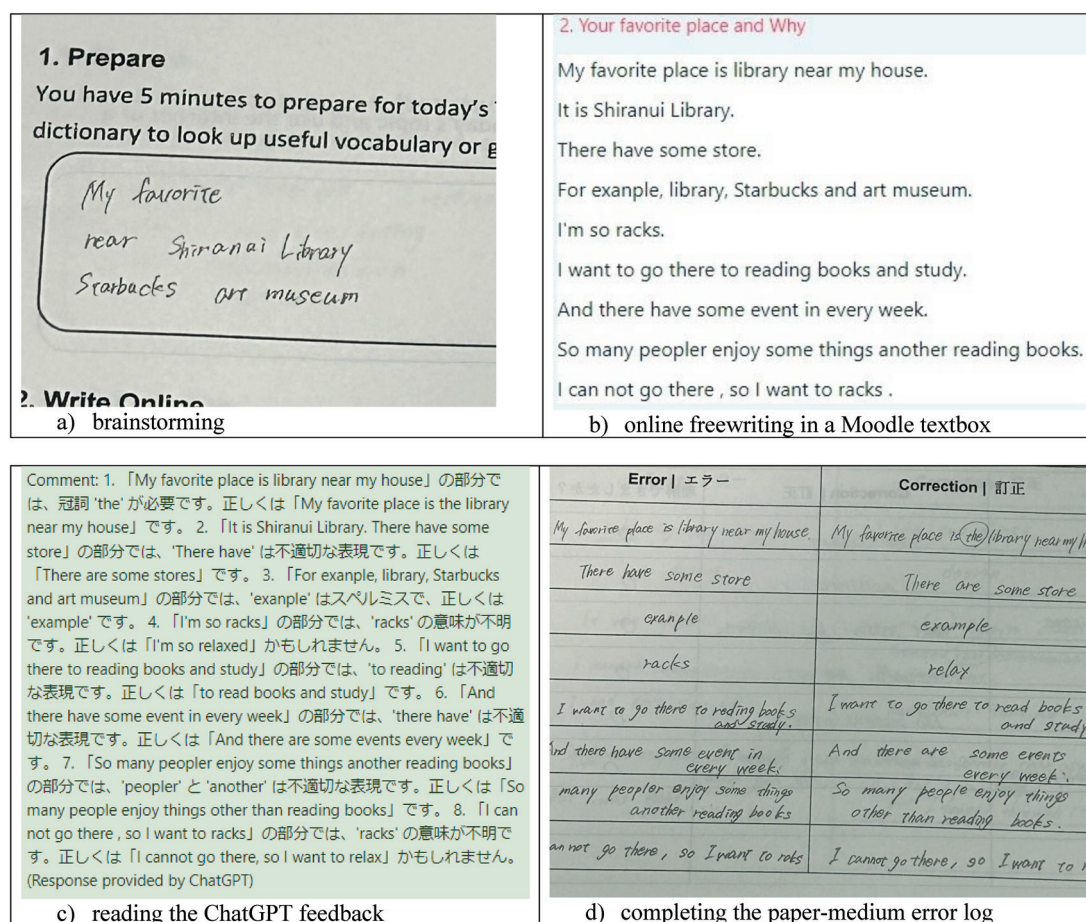


Figure 1 Classroom Intervention.

Once the intervention was begun and the students were completing the activity online, the teacher ceased to give comments on content and the only feedback was generated by ChatGPT, later reviewed by the researchers.

At the beginning of the first online activity, students were all given a 5-page booklet of free-writing activity guides (one page for each activity) with simple directions for brainstorming and writing down key vocabulary, simple instructions for writing online, a place to record their errors as indicated by ChatGPT, and two five-point Likert-scale items to complete, along with a space for optional comments. See Appendix A for the first activity guide.

The first step, *brainstorming*, comprised individual or collective brainstorming and an online search for potentially useful vocabulary for approximately five minutes. In this warmup to the main task, students were given the writing prompt (one of *a trip taken*, *last weekend*, *favorite season*, *student choice*, *summer plans*) and instructed to brainstorm useful vocabulary and phrases for the topic, to use electronic media, their classmates or the teacher as a resource for unknown English vocabulary, and to record the useful vocabulary in the first section of the paper activity guide. Students were also encouraged to discuss the topic in either English or their native Japanese as a kind of collective brainstorm. In practice, however, the vast majority of students were observed working solitarily at their computers (or, less commonly, with smartphones) looking up English vocabulary. The rationale for this brainstorming segment of the activity, frequently explained by the classroom teacher to students in Japanese, was to prepare for the writing segment of the activity, during which students were strongly discouraged from using any translation tools.

After five minutes, students began the subsequent *free-writing* component in an online Moodle textbox with copy and paste functions disabled by the teacher within the Moodle LMS. There was no access to a grammar or spell checker. Students were instructed to write in the Moodle textbox without referring to anything but their notes on the paper guide, to continue for a full 10 minutes, and to complete their final sentence after the 10 minutes had elapsed. In the several instances where students were observed with two windows open side by side, one open to an online translation tool, these students were reminded that this defeated the purpose of the activity and were asked not to use translation tools. Predictably, as will be described later, not all students were caught.

Following their writing, students clicked a ‘*check*’ button and waited between 15 and 30 seconds before *reading the ChatGPT feedback* in their L1 Japanese. At this time, the teacher walked about the classroom confirming that appropriate feedback had been generated, reminding students that ChatGPT feedback can be fallible, and encouraging students to ask questions.

The final step saw students *completing paper-medium error logs* (in the paper activity guide) to process their feedback. Students were asked to record in the leftmost column each of the up to 10 errors flagged by ChatGPT, provide a correction in the middle column, and indicate whether they understood the nature of each error in the third and final column. Following completion of the error logs, students answered two five-point Likert-scale items assessing the usefulness and comprehensibility of the feedback and had the option of providing any comments. At the end of the approximately 20–25 minute free-writing activity, the teacher collected the booklet. Appendix A displays the paper page for the first activity.

Teacher Prompt Engineering

Though Green’s (2024) plugin made it easy for students to receive feedback from ChatGPT, the onus of creating the prompt fell to the teacher. During the course of this study, the authors experimented

with different prompts for ChatGPT. Students always saw the *Question text* which necessarily varied for each assignment, but never saw the *AI Prompt* which directed ChatGPT in providing feedback (please see Figure 2). The classroom teacher began this study with a simple AI prompt, entirely in English, and then experimented with more complex prompts that included Japanese metalanguage and referenced example texts and feedback. Examples used in subsequent AI prompts were drawn from the ChatGPT feedback elicited by previous free-writing assignments.

For example, during the first round (Activity #1) of the online free-writing, one student's submission of *...I cannot study because I play game and watch TV...* elicited the feedback *... 'game' の前に冠詞 'the' が必要です [before 'game' the article 'the' is necessary]...* thus recommending *play the game* as a correction. The authors felt that the plural *play games* was more appropriate and we therefore revised the sample feedback in the ChatGPT prompt to read *"game" は単数形ですが、"play games" と複数形を使うことが一般的です* [*"game" is singular but it is typical to use the plural "play games" here.*]. In addition to revising the suggested corrections by ChatGPT, the authors also attempted to streamline the presentation of feedback that was not always consistent. Where the authors were satisfied with the earlier ChatGPT feedback, it was left unchanged in the model feedback sent as a part of the AI prompt.

There was concern, however, that time and effort spent creating modified AI prompts with numerous examples running several pages may not have resulted in better feedback. It was thus decided to experiment with two different AI prompts featured in Figure 2, the most complex for Activity #3 (shown in full in Appendix B) and the simplest for Activity #4, below:

You will be provided with English sentences. Your job is to find up to ten of the most serious grammar and spelling errors, explain the errors in simple Japanese and provide possible corrections.

The AI prompt for Activity #3 included Japanese metalanguage for terms such as tenses, transitive and intransitive verbs, subject-verb conjugation and others. The prompt also asked ChatGPT to reference four models of previous student writing along with the partially revised feedback as explained.

Question text ⓘ 2. Your favorite place and Why Words:6 AI Prompt ⓘ You will be provided with English sentences. Your job is to find up to ten of the most serious grammar and spelling errors, explain the errors in simple Japanese and provide possible corrections. Please explain using terms such as 過去形、現在形、未来形、現在完了形、品詞、自動詞、他動詞 and 受動態. Please also explain errors such as 三人称単数形など、前置詞、複数形、単数形、冠詞、代名詞、関係詞、接続詞. Please do not address matters of punctuation and spacing. Please use Examples 1-4 and the feedback for Examples 1-4 as a model. Example 1 I often study at home. However almost time I cannot study because I play game and watch TV . When I was high school, I often used study in the	Question text ⓘ What are you most looking forward to this summer? Words:9 AI Prompt ⓘ You will be provided with English sentences. Your job is to find up to ten of the most serious grammar and spelling errors, explain the errors in simple Japanese and provide possible corrections.
#3: long AI prompt with metalanguage and examples	#4: short AI prompt with no examples

Figure 2 Teacher Interface for AI Text Quiz Item: Long AI Prompt and Short AI Prompt.

Data Collection

Data were collected from three sources: post-activity evaluations, an end-of-semester evaluation, and feedback analysis of two subsets of writing entries and corresponding ChatGPT feedback.

Student Perceptions from the Post-activity Evaluation

Following completion of each activity, students were asked to respond to two five-point Likert-scale items soliciting an evaluation of usefulness and comprehensibility of the ChatGPT-generated feedback. There was also an optional space for any comments. The items were at the bottom of each activity guide as shown in Appendix A. Analysis of the post-activity evaluation data required Scheirer-Ray-Hare tests due to non-normal distributions.

Student Perceptions from the End-of-course Survey

For each of the three courses, on the final day of class the students were assigned to do the last of the free-writing activities and were subsequently asked to complete an online questionnaire (see Appendix C). The online questionnaire was similar to the post-activity evaluation with an added question about whether students were able to learn from the mistakes highlighted in the ChatGPT feedback. There were three other differences: the Likert-scale was modified to six points in order to avoid a “neutral” answer (Kusmaryono et al., 2022), the form was online, and it could be administered anonymously as no other student work accompanied the survey. Sixty-six students from the three classes completed the survey. For the Art and Engineering classes, there were two and three students, respectively, who were either absent or did not follow directions, resulting in response rates of approximately 90%. For the BioLife class, however, 13 students failed to respond to the survey, yielding a response rate of just over 50%. Due to non-normal distributions for means from three different classes, Kruskal-Wallis tests were used for analysis.

Feedback Data

The online free-writing activities described above were conducted five times each in two of the classes (Art, Engineering Faculties) and four times in the BioLife class. Writing and feedback data to answer the research questions was compiled from students’ online writing and corresponding ChatGPT feedback, all of which was automatically saved in the Moodle LMS and continually available online to both students and researchers.

The Activity #3 and Activity #4 writing and feedback for five students from the Art class was selected for analysis. The students were not chosen at random as, despite the teacher urging students to write without translation software and emphasizing that students got full credit simply for writing anything, and despite the disabling of the paste function in the LMS activity, some students had still undoubtedly used translation tools as evidenced by a clear mismatch in the writing on this assignment and other work. The first five Activity #3 entries that were clearly written by students were chosen for analysis. Unfortunately, Student O appeared to have used machine translation for Activity #4 and so this work was not further examined.

The authors together analysed the feedback for the Activity #3 and #4 writings from the four students. The first round of analysis was of successful ChatGPT corrections versus inaccurate or overlooked corrections alongside notes. Errors were labelled by type and counted, employing methodology and error codes adapted from Ferris (2006) and Hirschel (2011), and are included in Appendix D. This first round of error analysis was conducted on two separate days by both researchers coding together and

Table 1 Feedback Analysis of Participant R for Activity #3, First Four Errors Only

P	Error Code	Feedback	Fails	Com	Ins	Inc
R	art	1. 「My favorite place is library near my house」の部分では、冠詞 'the' が必要です。正しくは「My favorite place is the library near my house」です。		1		
R	wc	2. 「It is Shiranui Library. There have some store」の部分では、'There have' は不適切な表現です。正しくは「There are some stores」です。			1	
R	s/pl	2. 「It is Shiranui Library. There have some store」の部分では、'There have' は不適切な表現です。正しくは「There are some stores」です。			1	
R	sp	3. 「For example, library, Starbucks and art museum」の部分では、'exanple' はスペルミスで、正しくは 'example' です。	the library, an art museum	1		

Note. art = article, wc = word choice, s/pl = singular/plural, sp = spelling

making an initial analysis together on day one, and then reviewing and revising the analysis as necessary on day two. Nine additional errors were recorded on the second day. Table 1 above shows the feedback analysis of the first four errors of Participant R for Activity #3.

The second round of analysis examined the perceived comprehensibility of the feedback. The researchers together examined each instance of ChatGPT-generated feedback and jointly scored it as either fully comprehensible (Com), comprehensible but insufficient in explanation (Ins), or incomprehensible (Inc) as shown in Table 1 above.

There is subjectivity in what constitutes comprehensibility that we will explore further in the next section.

Results

Student Perceptions from the Post-activity Evaluation

Data from the post-activity evaluation (see bottom of Appendix A) from two classes (papers were inadvertently returned to the Art class) indicated that students generally appreciated the ChatGPT feedback. Descriptive statistics for activities 1 and 4 are shown in Table 2 below. For the four activities in the BioLife Class, students' averaged evaluation of the *usefulness* of the activity ranged from 4.43 for the first activity ($N = 21$, $SD = 0.75$) to 4.65 for the fourth activity ($N = 20$, $SD = 0.49$). Regarding *clarity* of the feedback, student's average evaluations ranged from 4.24 for the first activity ($N = 21$, $SD = 0.77$) to 4.65 for the fourth activity ($N = 20$, $SD = 0.49$). For the engineering classes, the averages were similar but appeared to decrease slightly between the first and fourth activities. The engineering students' averaged evaluation of the *usefulness* of the activity ranged from 4.60 for the first activity ($N = 20$, $SD = 0.50$) to 4.47 for the fourth activity ($N = 17$, $SD = 0.62$). As concerns *clarity*, the engineering student's average evaluations ranged from 4.50 for the first activity ($N = 20$, $SD = 0.76$) to 4.41 for the fourth activity ($N = 17$, $SD = 0.71$). Scheirer-Ray-Hare tests were performed and determined that there were no statistically significant differences in students' evaluations of usefulness and clarity between the first and fourth administrations of the activity and between the BioLife and Engineering

Table 2 *Descriptive Statistics for the First and Fourth Online Free-Writing Activities*

Free-writing Activity #1	N	M	SD	Free-writing Activity #4	N	M	SD
BioLife – Usefulness	21	4.43	0.75	BioLife – Usefulness	20	4.65	0.49
BioLife – Clarity	21	4.24	0.77	BioLife – Clarity	20	4.65	0.49
Engineering – Usefulness	20	4.60	0.50	Engineering – Usefulness	17	4.47	0.62
Engineering – Clarity	20	4.50	0.76	Engineering – Clarity	17	4.41	0.71

students. Thus, the slight uptick in ratings by BioLife students and the slight downtick for Engineering students can only be ascribed to chance.

Following the two Likert-scale items, there was a place for optional comments. Three students in the two classes wrote five comments, three of which indicated clear appreciation for the feedback, e.g. *It helps me notice small mistakes and grammatical errors*; one of which indicated that the student should be careful as she had made numerous spelling errors, and one comment which referred to the free-writing activity as challenging. Though there were several students in each class who sometimes failed to circle the Likert-scale items, opinions regarding *usefulness* and *clarity* consistently fell between the *Agree* and *Strongly Agree* descriptors for these post-activity items and the optional comments appeared to confirm a positive evaluation.

Student Perceptions from the End-of-course Survey

As shown in Table 3, the three classes indicated general satisfaction with the activity as evaluated by the end-of-course survey. The descriptive statistics in Table 3 show that the respondents from the Art ($n = 26$) and BioLife ($n = 14$) classes generally found the ChatGPT feedback to be useful (Art, $M = 5.35$, $SD = 0.94$; BioLife, $M = 5.36$, $SD = 0.93$), to be clear (Art, $M = 5.19$, $SD = 0.94$; BioLife, $M = 5.00$, $SD = 1.11$), and to be valuable for learning (Art, $M = 5.23$, $SD = 0.95$; BioLife, $M = 5.36$, $SD = 0.93$), means that all fell in the top band of the Likert-scale, *strongly agree*.

On the contrary, students from the Engineering class ($n = 26$) indicated lesser appreciation for the feedback with measures of usefulness ($M = 4.65$, $SD = 1.913$), clarity ($M = 4.81$, $SD = 1.13$) and learning ($M = 4.73$, $SD = 1.15$) falling in the second band, *agree*. Figure 3 below shows the distribution plot for responses concerning whether the ChatGPT-generated feedback was perceived to be useful.

Kruskal-Wallis tests were performed and determined that differences in reports of usefulness of the activity in the Art, BioLife and Engineering classes were significant, $H(2, N = 66) = 6.472$, $p = .039$, $\eta^2 = .071$, classified as a medium effect size (Plonsky & Oswald, 2014; Norouzian & Plonsky,

Table 3 *Descriptive Statistics from the End-of-Course Survey*

Art	N	M	SD	BioLife	N	M	SD	Engineering	N	M	SD
Usefulness	26	5.35	0.94	Usefulness	14	5.36	0.93	Usefulness	26	4.65	1.13
Clarity	26	5.19	0.94	Clarity	14	5.00	1.11	Clarity	26	4.81	1.13
Learning	26	5.23	0.95	Learning	14	5.36	0.93	Learning	26	4.73	1.15

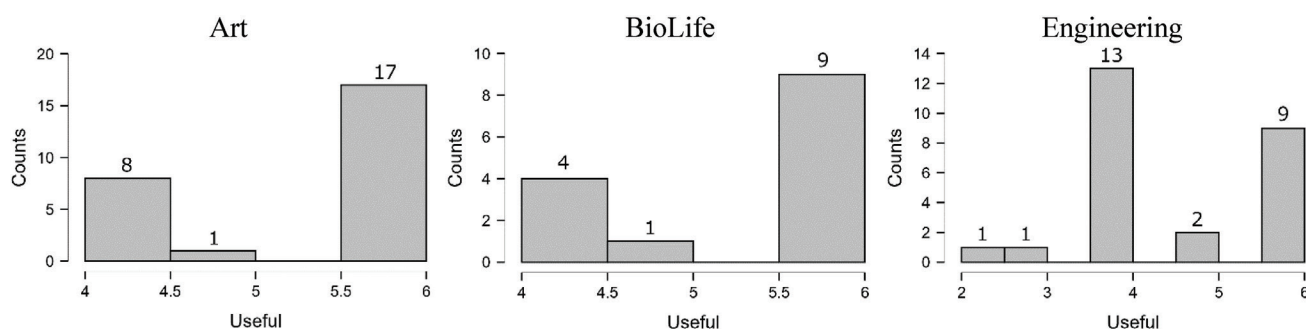


Figure 3 Frequency Distribution for Student Responses Regarding the Usefulness of ChatGPT Feedback.

2018). Post hoc Dunn tests revealed that Art students found the feedback significantly more useful than the Engineering students ($p = .021$), but that there were no statistically significant differences between the Art and BioLife students ($p = .985$), nor between the BioLife and Engineering students ($p = .052$). For the measure of clarity, a second Kruskal-Wallis test determined that there were no statistically significant differences between the three groups, $H(2, N = 66) = 1.447, p = .485, \eta^2 = 0$. Finally, for the measure of learning from the feedback, a third Kruskal-Wallis test determined that there were no statistically significant differences between the three groups, $H(2, N = 66) = 3.914, p = .141, \eta^2 = .060$.

In the open-ended comments section, there were 21 separate responses, all in Japanese. Twenty of the responses were clearly positive, referencing appreciation for the notations of mistakes, suggestions for corrections, opportunities for learning and/or a generally positive evaluation of the activity. Only one response could be construed as negative, *Writing things down doesn't stick in my memory that much*. This response likely refers to the error logs that students were asked to complete following the brainstorming, writing, and reviewal of feedback. This remark will be further explored in the Discussion.

Feedback Analysis

There were two components to the analysis of the feedback. The first was on accuracy and the second was on perceived comprehensibility for students.

Analysis of Feedback Accuracy

Analysis of feedback accuracy is perhaps best understood visually through Table 4 below. This table shows the feedback analysis for Participant R for Activity #3, whose submission was the following (edited only to remove paragraph breaks):

My favorite place is library near my house. It is Shiranui Library. There have some store. For example, library, Starbucks and art museum. I'm so racks. I want to go there to reading books and study. And there have some event in every week. So many peopler enjoy some things another reading books. I can not go there , so I want to racks.

As can be seen in Table 4, the researchers determined 15 instances where the ChatGPT feedback resulted in suggestions presumed to make the text more grammatically accurate, if not always eloquent. This determination, however, is clearly subjective. The researchers were unconcerned with sentence fragments such as *For example, the library, Starbucks and an art museum* nor were they troubled by a strange transition in the last sentence *I cannot go there, so I want to relax* which would arguably be better concluded with *...but I want to relax* or *...so I need to find another way to relax*. Without asking

Table 4 *Feedback Analysis of Participant R for Activity #3 in Full*

P	Error Code	Feedback	Fails	Com	Ins	Inc
R	art	1. 「My favorite place is library near my house」の部分では、冠詞 'the' が必要です。正しくは「My favorite place is the library near my house」です。		1		
R	wc	2. 「It is Shiranui Library. There have some store」の部分では、'There have' は不適切な表現です。正しくは「There are some stores」です。			1	
R	s/pl	2. 「It is Shiranui Library. There have some store」の部分では、'There have' は不適切な表現です。正しくは「There are some stores」です。			1	
R	sp	3. 「For example, library, Starbucks and art museum」の部分では、'exanple' はスペルミスで、正しくは 'example' です。	the library, an art museum	1		
R	sp	4. 「I'm so racks」の部分では、'racks' の意味が不明です。正しくは「I'm so relaxed」かもしれません。		1		
R	wf	5. 「I want to go there to reading books and study」の部分では、'to reading' は不適切な表現です。正しくは「to read books and study」です。			1	
R	wc	6. 「And there have some event in every week」の部分では、'there have' は不適切な表現です。正しくは「And there are some events every week」です。			1	
R	s/pl	6. 「And there have some event in every week」の部分では、'there have' は不適切な表現です。正しくは「And there are some events every week」です。			1	
R	prep del	6. 「And there have some event in every week」の部分では、' there have' は不適切な表現です。正しくは「And there are some events every week」です。			1	
R	sp	7. 「So many peopler enjoy some things another reading books」の部分では、'peopler' と 'another' は不適切な表現です。正しくは「So many people enjoy things other than reading books」です。			1	
R	word del	7. 「So many peopler enjoy some things another reading books」の部分では、'peopler' と 'another' は不適切な表現です。正しくは「So many people enjoy things other than reading books」です。			1	
R	wc	7. 「So many peopler enjoy some things another reading books」の部分では、'peopler' と 'another' は不適切な表現です。正しくは「So many people enjoy things other than reading books」です。			1	
R	word ins	7. 「So many peopler enjoy some things another ^ reading books」の部分では、'peopler' と 'another' は不適切な表現です。正しくは「So many people enjoy things other than reading books」です。			1	
R	sp	8. 「I can not go there , so I want to racks」の部分では、'racks' の意味が不明です。正しくは「I cannot go there, so I want to relax」かもしれません。		1		
R	sp	8. 「I can not go there , so I want to racks」の部分では、'racks' の意味が不明です。正しくは「I cannot go there, so I want to relax」かもしれません。			1	

Note. Com = fully comprehensible, Ins = comprehensible but insufficient in explanation, Inc = Incomprehensible, art = article, wc = word choice, s/pl = singular/plural, sp = spelling, wf = word form, prep = preposition, del = deletion, ins = insertion

the author, however, we can only guess what the student meant. This is where the *dialogic nature* afforded by the regular ChatGPT interface could be useful in helping users to co-construct meaning (Koltovskaia et al., 2024). Unfortunately, as of this writing, the setup of the AI Text tool allowed only for a single ChatGPT interaction per assignment via the OpenAI API.

After diagnosing the errors caught by ChatGPT, the researchers looked for what they termed *fails*. Again, there is some subjectivity here. The two fails noted in R's Activity #3 paragraph were articles preceding *library* and *art museum*. Some readers might also take issue with *It is Shiranui Library*, preferring to include the definite article: *It is the Shiranui Library*. As a proper noun, the researchers were satisfied without the article. The final sentence *I cannot go there, so I want to relax* is another place where one could reasonably argue that it is incoherent and should be treated as a fail. Given that grammaticality was the focus, we chose not to categorize this as a fail.

After counting the number of successful ChatGPT suggestions alongside the number of fails, the researchers determined rough accuracy ratings for Activities #3 and #4. Activity #3 with the long prompt appeared to yield a higher proportion of appropriate corrections and fewer inappropriate or overlooked corrections (82% accuracy over 62 errors) than the short prompt for Activity #4 (74% accuracy over 58 errors). However, as noted above, the counting of errors is notoriously difficult (Bryant et al., 2023) and so the numbers should be treated with caution. Additionally, the sample size ($N = 4$) was small.

It is also important to mention here that our analysis should not be mistaken for a comparison of ChatGPT-generated feedback vs. actual teacher feedback or even idealized feedback. Other studies have already focused on this strand of research and have found the feedback to be of sufficient quality (Kobayashi et al., 2024), and closer in alignment to instructor feedback than commercial applications such as Grammarly (Mizumoto et al., 2024).

Analysis of Feedback Comprehensibility

As with accuracy, comprehensibility is subjective and necessarily varies by person, hence the reason students were asked for their input in the two surveys. The researchers chose to evaluate comprehensibility by their perception of what a student at about an A1 level of the CEFRA with basic knowledge of metalanguage in Japanese could reasonably understand. By this reckoning, the feedback for student R's first sentence above 「*My favorite place is library near my house*」の部分では、冠詞 '*the*' が必要です [*in the section* 「...」 the article '*the*' is necessary] was deemed comprehensible as the missing article was identified and the correction after showed the correct placement. For more advanced learners, a teacher might additionally want to explain why the definite article is preferable (reference to a specific library), but we feel that for students at approximately the A1 level, this supplemental information may be too cognitively demanding. Indeed, some teachers may feel that employing meta-language is too demanding, preferring only to include corrections.

By contrast, the word choice correction of *There have some store* with '*There have*' は不適切な表現です [...*is an inappropriate expression*] was deemed comprehensible but insufficient. For our students (and probably any students who make such errors), a comprehensive explanation such as the following retrieved by the authors from ChatGPT (OpenAI, 2024) is clearly overwhelming as regards cognitive load (McFarlane & Spring, 2024), even if offered in the students' L1:

The verb “are” is a form of “to be,” which is appropriate for describing existence or location. The phrase “there have” is not used to indicate existence or presence. Instead, “have” is generally a verb associated with possession, obligation, or auxiliary functions (e.g., forming perfect tenses).

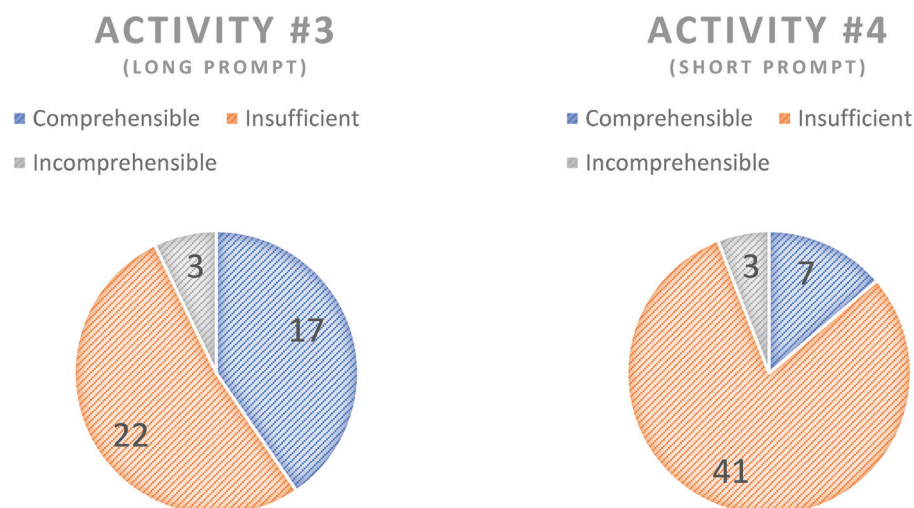


Figure 4 *Comprehensibility of ChatGPT-generated Feedback for Activities #3 and #4.*

Instead, an explanation approximating ここで正しい動詞は ‘are’ です [the correct verb here is ‘are’] was deemed preferable. In many instances of feedback, such as the pluralization of *store*, the error was corrected by ChatGPT in the 正しくは「…」 [The correct expression is...] portion of the feedback with no other mention nor explanation of the error. In all such instances the error corrections were labelled comprehensible but insufficient. It might not be surprising to the reader that corrections most frequently labelled comprehensible (and sufficient) were errors in spelling (11 of 25 total comprehensible corrections). The reason is that spelling errors, with the exception of alternative spellings in different varieties of English, are generally easy to identify and explain. Figure 4 above shows how the researchers categorized the ChatGPT-generated feedback for the two activities.

The long prompt for Activity #3 was more consistent in the format and language used to highlight errors and consistent in the language used to present the suggested corrections. Also, as desired, the long prompt resulted in more feedback that referenced, in Japanese, the type of error, such as misspelling, incorrect pronoun, wrong article, or inaccurate subject-verb conjugation. The seven comprehensible explanations generated by the short prompt were for misspellings (3), deletions (2), insertion of a determiner (1), and punctuation (1).

Discussion

Did Students Find the Feedback Meaningful?

As indicated by student responses to the post-activity evaluation and the end-of-semester survey, the students appeared to appreciate the grammatical feedback provided by ChatGPT. In this regard, the innovation was very much successful. Art and BioLife students indicated greater satisfaction with feedback on all measures (usefulness, clarity, facilitating learning) compared with Engineering students, the measure of usefulness between the Art and Engineering students achieving statistical significance. This difference could be attributable to a greater level of apathy among some Engineering students, as observed by measurably lower grades, inferior attendance, and a regular need for the teachers in this department to redirect some students back to their English study as opposed to furtively watching YouTube videos, playing games, or doing assignments for other classes. Such redirection was seldom needed in other classes. The question that necessarily arises is whether the feedback, and indeed the entire free-writing activity, is appropriate for all students. We submit that both the activity and the feedback are worthwhile, but that accommodations may need to be made, certainly on

a class level, but maybe even on an individual level. As concerns accommodations for the activity, for those students who insist on using machine translation, we may need to acknowledge that writing one paragraph in English unaided might be too challenging, too intimidating, or simply beyond a student's willingness to persevere. It may be necessary to furnish such students with modified assignments that help scaffold the activity by providing, for example, a bilingual list of possible topics, appropriate vocabulary and model sentences. Such a modified activity, while not necessarily ideal for the intended purpose of facilitating longer, independently composed texts, is certainly preferable to students writing in Japanese in a separate browser and uncritically typing the English output into the Moodle textbox. As concerns feedback, learner proficiency and engagement levels may more appropriately determine how many and which corrections are made, whether corrections are explained, and the level of detail provided in explanations (Liang et al., 2023).

From the optional comments in both the post-activity evaluations and the end-of-course survey, there was only one that we construed as possibly negative, and it was instructive: *Writing things down doesn't stick in my memory that much*. This point is important to consider and reflects the insistence by researchers (e.g., Chandler, 2003; Ferris, 2002, 2004; Hirschel, 2011) that merely making errors salient to students is not sufficient for learning. Something has to be done with that feedback. Traditionally, in writing classes, this something is the writing of multiple drafts. In the context of the current study, the error logs were a start, but clearly were not enough. One possible solution is the implementation of quizzes that have been individualized to reflect the most common grammatical errors of specific classes, or even perhaps individual learners. Students would be instructed to review their errors in preparation for quizzes that could conceivably be generated by AI. The potential for this sort of individualized instruction is part of what makes AI so appealing.

In summary, the answer to the first research question is that students do find the feedback meaningful, but that both the activity and feedback may need to be customized for each individual learner.

Did the Classroom Teacher Find the Feedback Meaningful?

As with the students, the classroom teacher was generally pleased with the grammatical feedback provided by ChatGPT. A non-specialist, non-AWF technology demonstrated a considerably high level of grammatical accuracy. Nevertheless, even the occasional unnecessary correction of a learner's perfectly natural English, an improperly corrected error, or an overlooked error can be demotivating for students and can cause them to lose faith in the feedback tool (Koltovskaia et al., 2024). Particularly for students who are unlikely to question the feedback, it is necessary to achieve a higher rate of accuracy, something that at least meets the highest levels reported by Shadiev and Feng (2024). Ideally, the chatbot would be able to recognize its limitations and flag a potential error as something that should be followed up with a teacher. For this reason, the authors have sought funding to build a finely-tuned model that can successfully diagnose and correct a wider variety and higher proportion of errors.

With regard to the comprehensibility of the feedback, the authors were also generally satisfied. We believe that the evolution of the prompts has led to clearer diagnosing of errors and a better presentation of possible solutions. Nevertheless, there is room for improvement. Numerous errors, despite being properly corrected, are still being labelled as 不適切 (incorrect) as opposed to having a simple grammatical explanation provided. Certainly, there are complicated errors for which an *incorrect* labelling and proposed correction is best. However, for simple and easily diagnosed errors, the authors would like the students to at least have access to why their writing is problematic. Such an example might return the following: *In "There are many scenery spots" the adjectival form of "scenery" is necessary. "There are many scenic spots" is correct*. Once again, with a finely-tuned model, we should be able to make such explanations in comprehensible Japanese for a wider variety of errors.

In summary, the answer to the second research question is that the classroom teacher does find the feedback meaningful, both in regard to accuracy and clarity. Congruent to the answer to RQ1 on student perceptions, we feel that the feedback can be improved through recommendations described below and summarized in the Implications section.

Concerns

Uncritical Acceptance of ChatGPT Feedback by Students

The authors are concerned that students took most or all feedback at face value, despite being warned of the fallibility of AI results. For example, the third column on the paper error logs asked students if they understood the nature of the reported error and/or feedback. Among the hundreds of errors recorded by students, not one student indicated on paper any uncertainty about an error. As discussed earlier, the authors found ChatGPT *reported errors* that comprised perfectly grammatical and natural English, as well as legitimate student errors that ChatGPT missed entirely. In some cases, the suggested *corrections* were grammatically suspect or unnatural, or did not reflect the intended locution of the author. There are several possible explanations for students failing to question the feedback. First, unlike the advanced graduate students in Koltovskaia et al.'s (2024) study, these students were comparatively low level and may not have had sufficient knowledge or confidence to question the feedback. Secondly, the students were all non-English majors taking required EFL courses, and their level of engagement did not always match expectations. Apathy was observed to be particularly acute in the faculty of engineering. This overall reduced level of engagement may explain the lesser appreciation for the activity by the engineering students. A similar explanation for students uncritically accepting feedback is that a lack of interest in English may have encouraged them to indicate understanding; doing otherwise invited a possibly unwelcome explanation from the teacher. A third possible explanation could be a cultural artifact whereby students were reticent to speak up (Harumi, 2011).

The Black Box, Prompting Issues, and Different Versions of ChatGPT

Of considerable concern is that ChatGPT, owned by for-profit company OpenAI, is not open-source. As such, there is a “black box” issue wherein we don’t know exactly how it works (OpenAI, 2025; Spring, 2024). We don’t know exactly what data it is trained upon as an unknown percentage is licensed and is not publicly disclosed (OpenAI, 2025). Yet Shi and Aryadoust (2024) have indicated that knowledge of the training data is necessary to ensure “compatibility between the educational goals of writing programs and AWF systems” (p. 202). Another unknown is how the system operates and how it processes prompts as, even if OpenAI were completely transparent about its inner workings, it is so complex that it is difficult even for experts in the field to understand. We don’t even know exactly how much of our prompts to ChatGPT are taken into consideration as the program has a *context window limit* of a certain number of tokens for both the prompt and the model’s output (OpenAI, 2025), a window that is not completely transparent. Furthermore, ChatGPT is continuously being updated and the performance is necessarily impacted by the model and its specifications (Mizumoto et al., 2024). Indeed, unbeknownst to the authors as they were executing this study, the model changed from GPT-4 to GPT-4o without warning or indication.

None of these “black box” issues is mentioned to suggest that ChatGPT should not necessarily be used as a source of AWF. These concerns are stated only to warn it is difficult to analyse how ChatGPT feedback is generated and that, even when researchers are transparent about models and prompts, it is impossible to guarantee consistency of performance and feedback as most researchers have no clear grasp of the workings of the chatbot at the center of everything.

Another concern in regard to prompting is that we prompted the chatbot to provide feedback on *ten of the most serious grammar and spelling errors*. There is no guarantee that ChatGPT interpreted *serious* in the same way we did. There is also the issue that we limited feedback to ten errors. The justification for this was not to overwhelm students with comprehensive feedback (Ferris, 2002) should there have been more than 10 errors in their paragraph of text. It is certainly possible, likely even, that our own prompt handicapped the chatbot's output.

Recommendations

Going forward with AI-generated CF, we have several suggestions to improve upon the intervention described in this paper.

Using an Open-source LLM

Transparency issues in regard to training data and how prompts are interpreted can be remedied by using open-source LLMs such as Ollama along with chatbots built specifically for integration with LMSs (Alier et al., 2025). Though Alier et al.'s Learning Assistant Manager and Builder (LAMB) requires some expertise and infrastructure to build an effective chatbot, as opposed to ChatGPT working immediately upon access, the researchers claim that it is possible for non-expert teachers to do, thus democratizing access to AI. Further, those authoring the chatbot will be able to incorporate the training data most applicable for their context, tailor the chatbot to respond appropriately to different prompts, and control the functionality and interface for users. With a custom built chatbot and an open-source LLM, it is even possible to provide feedback stating that the chatbot is unable to articulate the concern and that the student should consult a teacher (Alier et al., 2025). Such an acknowledgement of limitations would make the chatbot more reliable and could also prompt the chatbot's authors to input extra training data to remedy the shortcoming.

Selecting Errors to Correct

With increased functionality in AI Text and a custom-built chatbot, it should be possible for either the teacher or, ideally, the student to select errors for which feedback is generated, necessarily focusing the scope and preventing users from being overwhelmed (Ferris, 2002; McFarlane & Spring, 2024). Customization of individual learner feedback to certain error types is already available in some AWF programs (Shadiev & Feng, 2024).

Selecting Feedback Type

Improved functionality will also make it possible to enable the teacher or learners to select the type of feedback they wish to receive, direct feedback in which the learner is given suggested corrections, or indirect feedback in which the learner is made aware of the error and error type, but is not told explicitly what the correct form is. Indirect feedback is believed to offer a superior learning experience for motivated students (Liang et al., 2023), especially when the error type is easily diagnosed and easily fixed, such as for subject-verb agreement or singular-plural errors (Ferris, 2006). Ferris (2002) has noted, however, that the ideal feedback will vary by learner, suggesting that, if possible, the decision should be made by the learner, perhaps with guidance by the teacher.

Enabling Dialogue within AI Text

As of this writing, AI Text is configured to send only one query per quiz item to the connected chatbot via an API, failing to make use of the dialogic nature of chatbots. If the Moodle plugin can be mod-

ified to enable the learner several exchanges with the chatbot, the effectiveness of the tool could be greatly enhanced. It is important to remember, however, that the simplicity of the AI Text interface for the learner is part of its appeal and that there is necessarily a balance between functionality and usability.

Formatting

An important component of comprehensibility is formatting. While the current format is functional, it could be improved to be more user-friendly. One suggestion has been to provide inline formatting, whereby the error is flagged inside the text, better contextualizing it, much like the *Comments* function does in Microsoft Word. Another idea is the possibility of exploiting color screens/monitors such that we can highlight different types of errors in different colors (Shadiev & Feng, 2024). Improving the display of feedback is crucial in avoiding *cognitive overload*, a serious concern when performing a demanding task such as writing in a foreign language (McFarlane & Spring, 2024).

Limitations

This paper reports a small exploratory study of a new technology with just three classes of students, of which the assignments of only four students were carefully analysed. With larger sample sizes of students at different levels (the end-of-semester BioLife survey with 14 participants was particularly small), additional statistically significant results may be achieved. The students were approximately CEFR A1 level, non-English major students taking a required English course. Generalizability is therefore limited. Particularly in regard to higher level grammatical structures such as passive voice, reported speech and conditionals, there were few instances of use. As we seek to build a finely-tuned model, it is therefore imperative that we include training data from more diverse levels of students. Future studies should also necessarily aim for a participant population of varied level.

Single-item measures of the constructs usefulness, clarity, and learning opportunity are a further limitation. Multi-item scales (see, for example, Amrullah et al. 2023; Luo, 2023) might enable a more robust measure of student perceptions. Alternatively, richer qualitative data elicited from sources such as interviews or focus groups might give a fuller picture.

The use of ChatGPT as the AI chatbot is another limitation. The opacity of an otherwise extraordinary tool inherently means that its workings cannot be completely understood and cannot, therefore, be fully replicated. The change from GPT-4 to GPT-4o during the study may also have affected the consistency of feedback across different writing samples. A transition to an open source LLM and chatbot would enable greater clarity, replicability, and improvability.

Concerns of researcher bias must also be acknowledged. Analysis was completed by the authors, one of them the classroom teacher. Results should therefore be considered with caution. Future studies may wish to outsource analysis to peers or colleagues without a stake in the outcome and blinded to the nature of the texts vis-à-vis experimental group or control, first written submission or final, pre-test or post-test.

It should be noted here, that, as a study grounded in exploratory practice, the current investigation was necessarily limited to what was sustainable and could be integrated into the regular classroom practice (Allwright, 2005). Further, as a first foray into a new affordance in AI Text, an exploratory study with inherent limitations was believed to be the appropriate first step. Finally, despite the noted limitations, the authors nevertheless believe that the study has demonstrated successful use of a promising new application, the AI Text item type for Moodle.

Implications

Pedagogy

The biggest pedagogical implication is that generative AI combined with a plugin such as AI Text, can be used to provide accurate, comprehensible CF in a language learner's L1. The feedback can be accessed immediately following the writing task, when the content is still fresh in the learner's mind, as opposed to days later, after the teacher has marked it. Instead of spending considerable time and energy trying to diagnose student errors, teachers can redirect that time towards assisting students in learning from their errors. Nevertheless, there is room for improvement and several suggestions have been made. The first is that use of an open-source LLM would ensure that the training data are relevant and that the chatbot's operations are understood so that functionality can be continually improved. Enabling learners and their teachers the choice of which errors to have corrected and what type of feedback is provided can enhance effectiveness and learner autonomy. Enabling multiple interactions with the chatbot can also improve the efficacy. Finally, improvements to the formatting of feedback may better assist learners in engaging with the feedback.

Research

There are several implications for research. Though there have been numerous studies investigating chatbots using LLMs that are not open-source (e.g., ChatGPT), fewer studies have focused on open-source LLMs. Further research should investigate these possibilities as transparency in training data and chatbot operations could significantly improve the effectiveness of the feedback. If AWF systems are configured to provide both direct and indirect feedback (as in Liang et al., 2023), then additional research could more easily measure the efficacy of those conditions with different populations of learners (e.g. proficiency level, engagement level) across different types of errors. Similarly, with the capability of addressing only certain categories of errors, as opposed to all, research could be conducted to determine what categories of errors are best treated with what type of feedback. Finally, if the AWF systems enable students the choice of feedback type and feedback focus (which errors), research could investigate the effects on students in regard to engagement, efficacy, autonomy and motivation.

Conclusion

This paper has reported on the successful use of a ChatGPT-powered plugin to enable automated corrective feedback in the students' L1 immediately after students have written a text. Though there exists room for improvement in accuracy and presentation, students and the classroom teacher have found the feedback to be both comprehensible and useful for learning. Generative AI prompts of increasing complexity have been trialled, but it is believed that a finely-tuned model of an open-source LLM is necessary for further improvements in accuracy, comprehensibility, and effectiveness. Generalizability of this study is limited due to the small sample of participants at a comparatively lower level of language proficiency. Future research might examine different methods of applying LLMs to provide corrective feedback with a more diverse population of students.

References

- Alier, M., Pereira, J., García-Peñalvo, F. J., Casañ, M. J., & Cabré, J. (2025). LAMB: An open-source software framework to create artificial intelligence assistants deployed and integrated into learning management systems. *Computer Standards & Interfaces*, 92, 103940. <https://doi.org/10.1016/j.csi.2024.103940>

- Allwright, D. (2005). Developing principles for practitioner research: The case of Exploratory Practice. *The Modern Language Journal*, 89(3), 353–366. <https://doi.org/10.1111/j.1540-4781.2005.00310.x>
- Amrullah, A., Lail, H., & Sumayani, S. R. (2023). The EFL students' perspectives on the usefulness of ICT-based learning. *Journal of Language and Pragmatics Studies*, 2(1), 1–10. <https://doi.org/10.58881/jlps.v2i1.6>
- Bryant, C., Yuan, Z., Qorib, M. R., Cao, H., Tou Ng, H., & Briscoe, T. (2023). Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3), 643–701. https://doi.org/10.1162/coli_a_00478
- Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing*, 12(3), 267–296. [https://doi.org/10.1016/S1060-3743\(03\)00038-9](https://doi.org/10.1016/S1060-3743(03)00038-9)
- Ferris, D. R. (2002). *Treatment of error in second language student writing*. University of Michigan Press. <https://doi.org/10.3998/mpub.2173290>
- Ferris, D. R. (2004). The 'grammar correction' debate in L2 writing: Where are we, and where do we go from here? (and what to do in the meantime...?), *Journal of Second Language Writing*, 13(1), 49–62. <https://doi.org/10.1016/j.jslw.2004.04.005>
- Ferris, D. (2006). Does error feedback help student writers? New evidence on the short- and long-term effects of written error correction. In K. Hyland, & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 81–104). Cambridge University Press. <https://doi.org/10.1017/CBO9781139524742.007>
- Giessler, R. (2023). EFL writers' cognitive engagement with AWE feedback. *Language Awareness*, 33(2), 428–445. <https://doi.org/10.1080/09658416.2023.2269088>
- Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology*, 26(2), 5–24. <http://doi.org/10125/73474>
- Green, M. (2024). Moodle AI Text quiz question type by Marcus Green. GitHub. https://github.com/marcusgreen/moodle-qtype_aitext
- Harumi, S. (2011). Classroom silence: Voices from Japanese EFL learners. *ELT Journal*, 65(3), 260–269. <http://dx.doi.org/10.1093/elt/ccq046>
- Hirschel, R. (2011). A qualitative study in grammar logs. *International Journal of Pedagogies and Learning*, 6(2), 126–139. <https://doi.org/10.5172/ijpl.2011.126>
- Horai, K. (2025). Usage and perceptions of ChatGPT in class and self-access learning contexts. *SILC Journal*, 4(1), 22–32.
- Jiang, X., Li, J., & Chen, C. H. (2024). Enhancing critical thinking skills with ChatGPT-powered activities in Chinese language classrooms. *International Journal of Chinese Language Teaching*, 5(1), 47–73. <https://doi.org/10.46451/ijclt.20240105>
- Karataş, F., Abedi, F. Y., Ozek Gunyel, F., Karadeniz, D., & Kuzgun, Y. (2024). Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners. *Education and Information Technologies*, 29, 19343–19366. <https://doi.org/10.1007/s10639-024-12574-6>
- Kato, F. (2020). *Strategies for growing and enhancing university-level Japanese Programs*. Routledge. <https://doi.org/10.4324/9780429353451>
- Kirchmeyer, B. C. (2023). P-CHAT: Formative self-assessment using group oral discussion tasks. *Language Teacher*, 47(1), 38–41. <https://jalt-publications.org/sites/default/files/pdf-article/47.1-tlt-wired1-2.pdf>
- Kobayashi, M., Mita, M., & Komachi, M. (2024). Large language models are state-of-the-art evaluator for grammatical error correction. <https://arxiv.org/pdf/2403.17540>
- Koltovskaia, S., Rahmati, P., & Saeli, H. (2024). Graduate students' use of ChatGPT for academic text revision: Behavioral, cognitive, and affective engagement. *Journal of Second Language Writing*, 65, 101130. <https://doi.org/10.1016/j.jslw.2024.101130>

- Kusmaryono, I., Wijayanti, D., & Maharani, H. R. (2022). Number of response options, reliability, validity, and potential bias in the use of the Likert scale education and social science research: A literature review. *International Journal of Educational Methodology*, 8(4), 625–637. <https://doi.org/10.12973/ijem.8.4.625>
- Liang, K. H., Davidson, S., Yuan, X., Panditharatne, S., Chen, C. Y., Shea, R., Pham, D., Tan, Y., Voss, E., Fryer, L., & Yu, Z. (2023, July). ChatBack: Investigating methods of providing grammatical error feedback in a GUI-based language learning chatbot. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications* (BEA 2023) (pp. 83–99). <https://doi.org/10.18653/v1/2023.bea-1.7>
- Liu, G., & Ma, C. (2024). Measuring EFL learners' use of ChatGPT in informal digital learning of English based on the technology acceptance model. *Innovation in Language Learning and Teaching*, 18(2), 125–138. <https://doi.org/10.1080/17501229.2023.2240316>
- Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie, & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). Academic Press. <https://doi.org/10.1016/b978-012589042-7/50015-3>
- Luo, Z. (2023). Determinants of the perceived usefulness (PU) in the context of using gamification for classroom-based ESL teaching: A scale development study. *Education and Information Technologies*, 28(4), 4741–4768. <https://doi.org/10.1007/s10639-022-11409-6>
- Mackey, A. (2006). Feedback, noticing and instructed second language learning. *Applied Linguistics*, 27(3), 405–430. <https://doi.org/10.1093/applin/ami051>
- Mizumoto, A., Shintani, N., Sasaki, M., & Teng, M. F. (2024). Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment, *Research Methods in Applied Linguistics*, 3(2), 100116. <https://doi.org/10.1016/j.rmal.2024.100116>
- Moneus, A. M., & Sahari, Y. (2024). Artificial intelligence and human translation: A contrastive study based on legal texts, *Heliyon*, 10(6). <https://doi.org/10.1016/j.heliyon.2024.e28106>
- Norouzian, R., & Plonsky, L. (2018). Eta-and partial eta-squared in L2 research: A cautionary review and guide to more appropriate usage. *Second Language Research*, 34(2), 257–271. <https://doi.org/10.1177/0267658316684904>
- OpenAI. (2024). ChatGPT (Dec 22 version) [Large language model]. <https://chat.openai.com/chat>
- OpenAI. (2025). ChatGPT (Jan 7 version) [Large language model]. <https://chat.openai.com/chat>
- Özçelik, N. P., & Ekşi, G. Y. (2024). Cultivating writing skills: The role of ChatGPT as a learning assistant—a case study. *Smart Learning Environments*, 11(1), 10. <https://doi.org/10.1186/s40561-024-00296-8>
- Plonsky, L., & Oswald, F. L., (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912. <https://doi.org/10.1111/lang.12079>
- Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11(2), 129–158. <https://doi.org/10.1093/applin/11.2.129>
- Shadiev, R., & Feng, Y. (2024). Using automated corrective feedback tools in language learning: A review study. *Interactive learning environments*, 32(6), 2538–2566. <https://doi.org/10.1080/10494820.2022.2153145>
- 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>
- Spring, R. (2024). Rage against the machine: An argument for caution when using ChatGPT in EFL and for increased human-AI integration. *Proceedings of the 2024 STEM International Conference*, South Korea, 37–40. https://researchmap.jp/7000010038/published_papers/48220122/attachment_file.pdf
- Stockwell, G. (2024). ChatGPT in language teaching and learning: Exploring the road we're travelling. *Technology in Language Teaching & Learning*, 6(1), 2273. <https://doi.org/10.29140/tltl.v6n1.2273>

- Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principles and practice in applied linguistics* (pp. 125–144). Oxford University Press.
- Swain, M. (2000). The output hypothesis and beyond: Mediating acquisition through collaborative dialogue. In J. Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 97–114). Oxford University Press.
- Swain, M., & Suzuki, W. (2008). Interaction, output, and communicative language learning. In B. Spolsky & M. F. Holt (Eds.), *The handbook of educational linguistics* (pp. 557–570). Blackwell. <https://doi.org/10.1002/9780470694138.ch39>
- Windah, A., Murtisari, E., & Isharyanti, N. (2024). Lower-proficiency EFL students' use of Grammarly in writing: Behavior, cognition, and affect. *The JALT CALL Journal*, 20(1), 1–25. <https://doi.org/10.29140/jaltcall.v20n1.1089>
- Wolf, M. K., & Oh, S. (2024). Investigating the types and use of feedback in middle-school English language learners' academic writing. *Language Learning & Technology*, 28(1), 1–25. <https://www.lltjournal.org/item/10125-73583/>
- Zeevy-Solovey, O. (2024). Comparing peer, ChatGPT, and teacher corrective feedback in EFL writing: Students' perceptions and preferences. *Technology in Language Teaching & Learning*, 6(3), 1482. <https://doi.org/10.29140/tltl.v6n3.1482>

Appendix A

Free-writing Activity Guide

Name: _____ Student Number: _____ Date: _____

Free Writing # 1 Topic: _____

1. Prepare

You have 5 minutes to prepare for today's Free Writing. Think about today's topic and use the internet or a dictionary to look up useful vocabulary or grammar points.

2. Write Online

3. Record your errors from ChatGPT

Error エラー	Correction 訂正	理解できましたか? <i>Did you understand?</i>

フィードバックが役に立ちましたか?
Was the feedback useful?

強く思う←**5 4 3 2 1**→強く思わない
Strongly Agree Strongly Disagree

フィードバックが分かりやすかったですか?
Was the feedback easy to understand?

強く思う←**5 4 3 2 1**→強く思わない
Strongly Agree Strongly Disagree

コメント:
Comment:

Appendix B

Activity #3 AI prompt

You will be provided with English sentences. Your job is to find up to ten of the most serious grammar and spelling errors, explain the errors in simple Japanese and provide possible corrections. Please explain using terms such as 過去形、現在形、未来形、現在完了形、品詞、自動詞、他動詞 and 受動態. Please also explain errors such as 三人称単数形など、前置詞、複数形、単数形、冠詞、代名詞、関係詞、接続詞. Please do not address matters of punctuation and spacing. Please use Examples 1-4 and the feedback for Examples 1-4 as a model.

Example 1

I often study at home. However almost time I cannot study because I play game and watch TV . When I was high school, I often used study in the food court. with my friend. My obsession is to study in places that are little noisy. I often study while listening music. I prefer studying with partner. This is because having partner helps me concentrate.

Example 1 Feedback

1. 「However almost time」の部分では、「almost time」は不適切な表現です。正しくは「almost all the time」です。
2. 「play game」の部分では、冠詞 «a» または «the» を必要とします。また、「game» は単数形ですが、「play games」と複数形を使うことが一般的です。
3. 「When I was high school」の部分では、前置詞 "in" が抜けています。正しくは「When I was in high school」です。
4. 「I often used study in the food court. with my friend」の部分では、「to」が抜けています。正しくは「used to」です。
5. 「little noisy」の部分では、冠詞 «a» を必要とします。正しくは「a little noisy」です。
6. 「while listening music」の部分では、「to」が抜けています。正しくは「while listening to music」です。
7. 「with partner」の部分では、冠詞 «a» を必要とします。正しくは「with a partner」です。
8. 「having partner helps me concentrate」の部分では、冠詞 «a» を必要とします。正しくは「having a partner helps me concentrate」です。

Example 2

I prefer studying with friends. Because we can teach each other. And I prefer studying in some noise. I don't like wuiet places because it makes me sleepy...
I feel like I'm not studying now. There are a lot of things that I want to study. So I think it's good to make a plan. I have a lot of times of studying now because I'm a student. So I think it's important to use my time effectively to study.

Example 2 Feedback

1. 「I prefer studying with friends. Because we can teach each other.」の文章では、接続詞を使って一つの文にまとめた方が良いです。正しくは「I prefer studying with friends because we can teach each other.」です。
2. 「studying in some noise」の部分では、前置詞 «in» の代わりに「with」を使います。正しくは「studying with some noise」です。
3. 「wuiet places」では、「wuiet」はスペルミスで、正しくは「quiet」です。

Example 3

My home is the best place to study. Because I like quiet place. I can't concentrate to study with a partner. I often study while listening to music. But I could not concentrate to study. So I think we should not listening to music while study. I don't like study. So I think I study not enough.

Example 3 Feedback

1. 「My home is the best place to study. Because I like quiet place.」の文章では、接続詞を使って一つの文にまとめた方が良いです。正しくは「My home is the best place to study because I like quiet places.」です。
2. 「I like quiet place」の部分では、一般的に「like」の目的語は複数形を使います。正しくは「I like quiet places」です。
3. 「concentrate to study」の部分では、»to study»の代わりに»on +ing»を使います。正しくは「concentrate on studying」です。
4. 「should not listening」の部分では、助動詞»should»の後には動詞原形が必要です。正しくは「should not listen」です。
5. 「while study」の部分では、正しくは「while I study」または「while studying」です。
6. 「I study not enough」の部分では、正しくは「I don't study enough.」です。

Example 4

My favorite season is spring. because the temperature is pleasant in spring. It is also a beautifull season when the cherry blossoms are in full bloom. On the other hand, I hate is the rainy season. because, I rainy season is hot and wet. also temperature is too hot in rainy season. but, I like rainy day. because I feel relax in rainy and I like rain smell.

Example 4 Feedback

1. 「My favorite season is spring. because the temperature is pleasant in spring.」の文章では、接続詞を使って一つの文にまとめた方が良いです。正しくは「My favorite season is spring because the temperature is pleasant in spring.」です。
2. 「beautifull season」では、「beautifull」はスペルミスで、正しくは「beautiful」です。
3. 「I hate is the rainy season」の部分では、不必要な「is」が含まれています。正しくは「I hate the rainy season」です。
4. 「I rainy season is hot and wet」の部分では、「I」が不適切です。正しくは「The rainy season is hot and wet」です。
5. 「also temperature is too hot in rainy season」の部分では、冠詞「the」が必要です。正しくは「Also, the temperature is too hot in the rainy season」です。
6. 「but, I like rainy day」の部分では、一般的に「like」の目的語は複数形を使います。正しくは「But I like rainy days because I feel relaxed in rainy weather」です。
7. 「because I feel relax in rainy」の部分では、形容詞「relaxed」を使用すべきです。また、「rainy」の後に「weather」が必要です。正しくは「because I feel relaxed in rainy weather」です。
8. 「I like rain smell」の部分では、冠詞「the」と関係詞「of」が必要です。正しくは「I like the smell of rain」です。

Appendix C

End of Semester Survey

Chat GPT (AI)を使ったフリーライティング

Using ChatGPT in free writing

* Required

1. 学部 * ☐ Faculty

☐ 工学部

☐ 芸術学部

☐ 生物生命学部

When using ChatGPT in free writing

2. Chat GPT (AI)を使ったフリーライティング * ☐

	strongly agree	agree	agree slightly	disagree slightly	disagree	strongly disagree
	非常に 賛成	賛成	少し 賛 成	少し 反 対	反対	非常に 反 対
...でのフィードバックが役に立った。 ...the feedback is useful.	☐	☐	☐	☐	☐	☐
...でのフィードバックが分かりやすかった。 ...the feedback is easy to understand.	☐	☐	☐	☐	☐	☐
...で指摘された間違えから学ぶことができた。 ...the feedback enables me to learn from my mistakes.	☐	☐	☐	☐	☐	☐

3. コメント ☐

Enter your answer

Submit

Appendix D

Error Codes

Code	Error Type	Example
wc	word choice	she did ^{wc} me good advice
wf	word form	my son is impatience ^{wf}
tense	tense	we make ^{tense} pasta last night
art	articles (a, an, the or ∅)	planets in ^{art} universe
art/det	article or determiner (this, my John's, etc.)	^{art/det} Friend wants to try sushi
art/pl	article, determiner, or plural	he brought gift ^{art/pl} for the children
prep	preposition	she is worried to ^{prep} you
s/pl	singular/plural	we played three game ^{s/pl}
agr	subject-verb agreement	his trousers is ^{agr} stylish
sp	spelling	that infomation ^{sp} is on the back
ins	insertion	Jun can speak languages other ^{ins} English
del	deletion	I bought to ^{del} some Gundam goods

Adapted from Ferris (2006) and Hirschel (2011)