16
Feedback and Automated Writing Evaluation (AWE)

Inyoung Na
Mahdi Duris
Volker Hegelheimer

Pre-reading questions:

1) How can *ChatGPT* benefit second language (L2) learners in receiving writing feedback?
2) What factors might influence learners' attitudes towards using *ChatGPT* as a tool for improving English writing skills?

Introduction

Automated Writing Evaluation (AWE) systems, such as Grammarly and ETS's Criterion, have gained considerable interest for integration into language classrooms to provide second language (L2) learners with corrective and diagnostic feedback. The practical benefits of these systems were recognized for alleviating the workload of instructors, enabling them to focus their feedback efforts on aspects of writing that require human evaluation, such as discourse features of writing, rather than on sentence-level issues like grammar, word choice, and mechanics (Ranalli et al., 2017). Students receiving immediate and frequent feedback through AWE tools showed enhanced L2 development, increasing autonomy and motivation to write (Warschauer & Grimes, 2008).

With the recent release of highly sophisticated large language models (LLMs), such as OpenAI's GPT series and Google's PaLM2, AWE seems to be entering a new era. These models are trained on vast amounts of text data from the internet to understand and predict

language patterns, generating coherent human-like text in response to user prompts. Simple chat interfaces like ChatGPT by OpenAI and, more recently, Google's Gemini and Microsoft's Copilot have allowed an increasing number of users to interface with LLMs to perform tasks as if they are interacting with a human. When this capability is used in L2 writing contexts, particularly for automated writing evaluation and feedback, LLMs may provide students with support like what they receive from teachers. This allows them to request feedback about various aspects of their writing, including global aspects such as organization and flow, in addition to the sentence-level error corrections offered by traditional AWE tools. Thus, we aimed to explore what interactions between L2 learners and LLMs look like and the extent to which LLMs offer the support that students need, potentially leading to revisions across different aspects of writing. Given that ChatGPT has been available the longest (since November 2022) and is the most widely used, we chose ChatGPT based on GPT-3.5 as the AI chat platform for learners to interact within our study.

To evaluate ChatGPT's effectiveness as an AWE tool, it is crucial to examine the interaction between the learner and ChatGPT due to its reactive nature. Specifically, the effectiveness of ChatGPT's feedback is largely determined by the user prompt, as it responds based on what is requested. Therefore, we investigated learner-ChatGPT interactions using Chapelle's (2001) CALL evaluation framework, adopting an interactionist approach to focus on ChatGPT's language learning potential. Additionally, it is essential to explore learners' detailed perspectives and attitudes towards ChatGPT, as these can provide insight into why students engage with ChatGPT in particular ways, such as accepting or ignoring its feedback.

However, it is important to emphasize that our intention is not solely to advocate for the benefits of using LLMs. A growing body of literature addresses concerns associated with LLMs in English writing, such as academic malpractice, inaccuracies, reduced learning, and misinformation (e.g., Fuchs, 2023; Song & Song, 2023). The findings of this study also highlight these issues, particularly in relation to plagiarism concerns. In what follows, the literature review will discuss AWE studies, LLMs in L2 Writing, and Chapelle's CALL Evaluation Framework.

Literature review

Automated writing evaluation

Automated Writing Evaluation (AWE) tools are digital writing environments that provide automated feedback (Cotos, 2023). AWE has been found to be more adept at providing feedback on sentence-level correctness than on higher-level concerns (Ranalli et al., 2017; Weigle,

2013), leading us to question whether ChatGPT is effective at one and/or the other. Moreover, learners' engagement levels and the subsequent effectiveness of AWE can be influenced by individual differences, highlighting the importance of researching students' variability, such as their learning orientations, proficiency, and perceptions of AWE's efficacy (Chen et al., 2022; Ranalli, 2021; Zhang & Hyland, 2018; Zhang, 2020), and through a qualitative and close lens (Godwin-Jones, 2022). For instance, even though students may revise their drafts using the AWE feedback, a closer look might tell us that they do not necessarily engage more deeply with and learn from the feedback than just adopting a quick proofreading orientation to the drafts (Stevenson & Phakiti, 2019), not going beyond sentence-level changes.

Learners perceiving ChatGPT as an effective writing tool are more likely to continue to use the AI tool and benefit from it further as a complement to the teacher's feedback and/or outside the language classroom.

LLMs and L2 writing

Recently, more studies have started to explore LLMs in the context of L2 writing, highlighting the benefits of ChatGPT's feedback enhancing students' skills in coherence, organization, vocabulary, grammar, and organization as well as students' positive perceptions of its use (e.g., Ali et al., 2023; Boudouaia et al., 2024; Mahaptara, 2024; Song & Song, 2023). For instance, Mahapatra (2024) evaluated ChatGPT's effectiveness in improving the writing skills of undergraduate ESL students in India, finding that those who received ChatGPT's feedback significantly outperformed a control group on writing tasks and expressed positive perceptions of its assistance with content, organization, and grammar. Similarly, Boudouaia et al. (2024) reported that EFL students in Algeria who used ChatGPT-4 for revising their texts over a 10-week intervention showed notable improvements in both local (e.g., grammar) and global (e.g., coherence) aspects of writing, compared to a control group that received traditional teacher feedback. Questionnaire results further revealed growing acceptance of ChatGPT's feedback among students, particularly regarding its perceived usefulness and ease of use. While it serves as an effective tool for writing practice and feedback that supplements teacher's feedback (Guo & Wang, 2023), some studies have also raised concerns about potential issues such as plagiarism, inaccuracies, over-reliance, and unoriginality in students' work (Fuchs, 2023; Kohnke & Moorhouse, 2023; Mahapatra, 2024; Xiao & Zhi, 2023).

However, the methodologies employed in these studies have been limited to experimental designs, intervention studies, and survey-based approaches. Many, including Mahapatra (2024), Boudouaia et al. (2024), and Song and Song (2023), used pre-test and post-test designs

to evaluate changes in writing proficiency, comparing experimental groups receiving ChatGPT feedback with control groups receiving traditional teacher feedback. Yet, this focus on test scores and group comparisons alone does not provide detailed insights into how students engaged with and used ChatGPT's feedback. As with previous AWE studies, a qualitative lens may suggest only a quick proofreading orientation using ChatGPT's feedback to the draft (Ranalli, 2021; Stevenson & Phakiti, 2019) rather than deep engagement with the writing process. More detailed explorations of the nature of learner-ChatGPT interaction are needed.

Chapelle's CALL evaluation framework

To evaluate ChatGPT's potential as an AWE system for promoting language learning, we applied Chapelle's (2001) Computer-Assisted Language Learning (CALL) evaluation framework. This framework provides six criteria for evaluation: language learning potential, meaning focus, learner fit, impact, authenticity, and practicality. Our study focuses on the evidence of language learning potential (LLP) through a detailed investigation of learners' patterns of using ChatGPT and their perceptions of its effectiveness. Here, LLP is broadly defined as the potential of feedback to facilitate a beneficial focus on form and discourse features (e.g., organization and development). Using the interactionist approach (Gass & Mackey, 2006; Long, 2007) to assess LLP, we examined learners' interactions with ChatGPT, focusing on how they attended to and noticed the feedback, as well as how they engaged with the modified input (i.e., learners' original output analyzed and returned with feedback by the system) to negotiate meaning and modify their output (see Figure 1 for this interaction process). It is important to note that since ChatGPT is reactive rather than proactive, the nature and quality of its feedback greatly depends on the types of feedback sought through user prompts. Therefore, our particular interest was to see how learners and ChatGPT co-construct the feedback and add another mechanism to maximize learning opportunities.

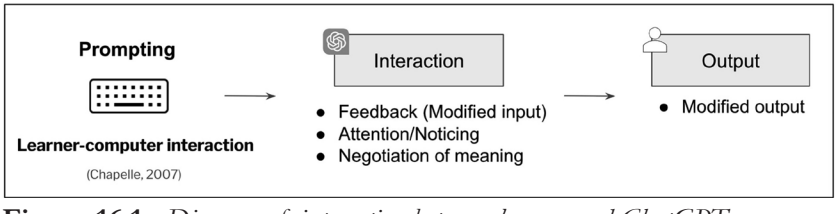

Figure 16.1. *Diagram of interaction between learner and ChatGPT.*

To the best of our knowledge, no prior study has rigorously examined how learners interact with ChatGPT for English writing revisions based on a theoretical framework such as Chapelle's CALL framework or the interactionist approach. Such detailed explorations will provide a better understanding of what aspects of the nature of learner-ChatGPT interaction shape learners' modified output–something that cannot be revealed through mere quantitative comparisons between experimental and control groups. Researching this topic is timely, considering the growing interdisciplinary interest in LLMs and attention to providing a detailed perspective from learners who currently or potentially use ChatGPT as an AWE tool. Insights from our study are also expected to help language teachers guide students to use ChatGPT more effectively with appropriate prompts and recognize its benefits for ongoing language development. Therefore, our exploratory study investigates the usage patterns and perceived effectiveness of ChatGPT, examining eight international students from two university-level ESL writing courses at different levels. Using student-ChatGPT conversation logs, semi-structured interviews, surveys, and screen recordings, we explored how these students interacted with and evolved their views on ChatGPT's support for their L2 writing across three revision sessions. The following research questions guided our study:

1. What interaction patterns are observed between the L2 learners and ChatGPT? What prompts do they use, what feedback do they receive, and how do they respond to this feedback?
2. What are the L2 learners' perceived effectiveness and attitudes towards ChatGPT in writing revisions and improvement?

Based on our findings from research questions 1 and 2, we evaluated the LLP of ChatGPT for improving English writing.

Methodology

Participants

The participants were eight international students (male = 6, female = 2) at a large university in the Midwestern U.S. They were recruited from a two-course ESL writing program, with five participants from the lower-level course and three participants from the higher-level course in that program. Participants completed background questionnaires through Qualtrics (https://qualtrics.com). All had experience using either the basic or plus version of ChatGPT except for participant B05. The characteristics of participants are summarized in Table 1.

Table 16.1 *Characteristics of participants*

ID	L1	Sex	Age	Student Level	TOEFL/IELTS/ etc. (converted to TOEFL score in parentheses)	Major	Type of ChatGPT user
B01	Arabic	Male	25	G	TOEFL 84	Community and regional planning	Weekly
B02	Telugu, Hindi	Female	25	G	IELTS 6.5 (79–93)	Information systems and business analytics	Weekly
B03	Telugu	Male	24	G	TOEFL 99	MBA	Daily
B04	Tamil	Male	23	G	IELTS 7 (94–101)	Industrial Design	Daily
B05	Portuguese	Male	29	G	Duolingo 105 (70–75)	Mechanical Engineering	Inactive
C06	Korean	Female	21	U	Duolingo 130 (98–103)	Statistics	Daily
C07	Bangla	Male	21	U	Cambridge-B (80–96)	Mechanical Engineering	Weekly
C08	Nepali	Male	18	U	SAT 590	Computer Engineering	Weekly

Note. B and C indicate whether the participant was from the lower writing course (B) or higher writing course (C); U=Undergraduate and G=Graduate.

Data collection

Four data collection sessions were conducted over three weeks: the first included a tutorial and was held in the first week, the second and third sessions in the second week, and the final session in the third week. Participants completed a revision task using ChatGPT during the first three sessions, followed by a brief Qualtrics survey about their experience. In the final session, participants engaged in semi-structured interviews along with stimulated recalls, individually with the researcher. Each session lasted between 45 and 60 minutes.

Tutorial and biography task

In the first session, participants watched a 5-minute tutorial on using ChatGPT to revise their essays. The tutorial provided an example of basic prompting, such as, "Could you give me general feedback on this essay?" and more specific prompting to direct the feedback toward

their needs and problems, including, "How can I make my thesis statement stronger?" "Can you provide feedback on the organization of my essay?" and "Are there any grammatical errors?" The same example prompts were also provided as written instructions. The purpose of the study was to observe how learners would interact with ChatGPT without being encouraged to use specific prompts. Therefore, no additional instructions were given beyond these examples, aside from emphasizing that ChatGPT's feedback is co-constructed through interaction with the user. This approach was taken to avoid influencing how participants used the tool and to prevent any potential skewing of the results. After watching the tutorial, participants were asked to write a short biography of 150–200 words in 20 minutes and then revise it using feedback obtained from ChatGPT in another 20 minutes. They submitted their work when ready and completed a short survey about their experience. This session lasted 50 minutes.

Revision task

In the subsequent two sessions, the participants completed a revision task using texts they had recently written or were in the process of writing for the course they were currently enrolled in. The decision to have participants bring their own authentic texts rather than assigning a task in a controlled setting was intended to encourage them to invest more fully in the revision process. Because they finished revising the texts earlier than expected, all participants revised one text per session, resulting in two texts each across the two sessions. Each participant worked individually on a provided computer in a private lab. The researcher was available for technical assistance but did not observe the tasks directly. The participants submitted their texts to ChatGPT-3.5, prompted it at their discretion to receive feedback, and revised their texts accordingly in a Google Docs document. They worked through the feedback until they were satisfied with their writing or reached the 40-minute time constraint. The participants submitted their completed drafts, and conversations between them and ChatGPT were automatically archived. After submission, they completed the same survey regarding their experience.

Survey questions

The survey included 10 Likert-scale questions regarding participants' general views on ChatGPT and its feedback, their perceived effectiveness, and attitudes and reactions towards using ChatGPT for English writing revision. Survey questions were adapted from Cotos (2010) and

Escalante et al. (2023). Each question was rated on a five-point scale, with one being the lowest (not at all or very poor) and five being the highest (very good or very much). Four questions focused on perceived effectiveness as an indirect measure of ChatGPT's potential for language learning, leading to four variables: the extent of changes learners made based on ChatGPT's feedback (change), the extent to which ChatGPT helped learners notice sentence-level errors (local) or global concerns such as content and organization (global), and the development of English writing skills overall (development). Another four questions focused on participants' attitudes towards using ChatGPT, including their satisfaction with ChatGPT's feedback (satisfaction), preference for using ChatGPT over teacher feedback (preference), trust in ChatGPT's feedback (trust), and willingness to use ChatGPT for future writing tasks (future).

Screen recording

The default screen recording function of the desktop was used to capture the data on participants' interaction with ChatGPT, producing video files that document all visible actions on their computer screens.

Semi-structured interviews/stimulated recalls

In the final session, semi-structured interviews were conducted with each participant individually. The participants were asked about how they interacted with ChatGPT and their perceived effectiveness and attitudes towards using ChatGPT for language learning purposes. To further contextualize their responses, participants were also asked about the text they had provided, their previous experiences with ChatGPT, and their experience with English writing and receiving written feedback.

After the interview, stimulated recalls were conducted. Individualized questions were prepared based on researchers' notes about interesting points in the participants' patterns of using ChatGPT, as observed in the screen recordings. Participants were shown these parts of the recordings and prompted to verbalize their thoughts as needed. The questions focused on why the participants addressed the feedback in a particular way, for example, "Why did you ignore this feedback?" and "Why did you type out ChatGPT's output rather than copying and pasting it into your draft?" Other questions aimed to determine whether the participants noticed the errors or changes made by ChatGPT and whether they learned from the feedback. Semi-structured interviews and stimulated recalls were all audio-recorded.

Data coding

To address RQ1, the student-ChatGPT conversation logs were annotated using MAXQDA 2024 (VERBI Software, 2024) to capture two variables:

- **Prompt Type**: What prompt learners used or the specific interaction they sought with ChatGPT (e.g., asking for feedback, revision).
- **Feedback Type**: What types of feedback or output ChatGPT produced in response to the prompts (e.g., summative comment, revision).

Response type, or how the learner reacted to ChatGPT's feedback, was identified through an analysis of screen recordings. We compiled all segments of prompts, outputs, and responses, created initial codes, and then recoded based on a refined list of codes.

To address RQ2, audio recordings of the stimulated recalls/semi-interviews were transcribed and imported into MAXQDA. The transcripts were analyzed following Corbin and Strauss (2007), which involved open coding (identifying relevant units and initial codes for analysis) and axial coding (making connections between codes to create categories). The categories identified during the axial coding stage were refined into a final set of eight codes: effectiveness/ineffectiveness of feedback, liking/disliking about the feedback, learning/no learning, learning to use ChatGPT, perception of ChatGPT, emotional reactions, trust/partial trust, preference and continued use. These codes were then used to recode the dataset.

Data analysis

To address RQ1, percentages were calculated for the occurrence of each type of prompt and feedback coded in the student-ChatGPT conversation logs, allowing for a quantitative analysis of patterns and trends in interactions with ChatGPT. Response types were identified through the screen recordings. To address RQ2, descriptive statistics for Likert-scale responses in the survey were calculated and compared. Qualitative analyses involved writing analytic memos for each of the eight participants based on the coded transcriptions, biodata from the interviews, and survey data analyzed for each participant.

Results

Interaction with ChatGPT: prompt, feedback, and response

RQ1 addressed the interaction patterns between the learner and ChatGPT, specifically focusing on learners' prompt types, ChatGPT's

Table 16.2 *Prompt types and descriptions*

Prompt Type	Description	Count (%)
Asking for Feedback	The participant asks for feedback on their text. *(e.g., provide specific feedback on the fourth paragraph; can you provide me some examples?)*	143 (45.25)
Revision Request	The participant asks ChatGPT to rewrite the text. *(e.g., make it better and polished in 200–250 words; rewrite the second paragraph with all your suggestions)*	75 (23.73)
*Follow-up	The participant follows up on ChatGPT's response with a further request or meaning negotiation. *(e.g., I don't feel that giving examples is a good point; with more simple language)*	73 (23.10)
Information Search	The participant searches for general information not related to the target language. *(e.g., how much CO2 is produced to make one car)*	12 (3.80)
Evaluation Request	The participant asks ChatGPT to grade their paper. *(e.g., Would you rate this out of 100?)*	7 (2.22)
Text Generation	The participant asks ChatGPT to generate new text without basing it on their own text. *(e.g., write your response on the findings)*	6 (1.90)
		Total: 316 (100%)

Note: Follow-up was double-coded with any other type.

feedback types, and learners' response types. We identified six prompt types (Table 2). Among the 316 prompts identified, 'asking for feedback' was used the most (45.25%), followed by 'revision request' (23.73%), and 'follow-up' with another prompt (23.10%). Among 'follow-up', there were 9 instances (12.33%) of meaning negotiation.

Regarding 'Feedback' (Table 3), among ChatGPT outputs, we identified nine methods for how ChatGPT addressed participants' prompts for a total of 282 coded segments. In most cases, prompt type and feedback type corresponded to each other. However, even in cases where the participant only asked for feedback, ChatGPT sometimes provided fully revised text. As shown in Table 3 below, 'Revision' was the most frequently used method (40.43%), followed by 'Suggestion' (26.60%) and 'Summative comment' (15.10%).

Lastly, through the analysis of screen recordings, we identified three 'Response' types (Table 4), or in what ways participants made use of ChatGPT's feedback, along with other interesting behaviors (Table 5) we observed. For 'No modified output', participants either ignored the

Table 16.3 *Feedback types and descriptions*

Feedback Type	Description	Count (%)
Revision	ChatGPT produces the revised text.	114 (40.43)
Suggestion	ChatGPT gives suggestions for improvement, usually providing a bulleted list of points. *(e.g., Vocabulary Enhancement: Introduce more diverse and sophisticated vocabulary where appropriate to add depth and nuance to the text.)*	75 (26.60)
Summative comment	ChatGPT provides a summary of evaluative comments on the writing. *(e.g., The biography is clear and well-structured.)*	44 (15.60)
Direct correction	ChatGPT indicates that there was an error in the writing with the correct form. *(e.g., Changed "comes" to "stems" for better tense agreement.)*	13 (4.61)
Information	ChatGPT provides general knowledge not related to the target language. *(e.g., Red symbolizes passion, energy, and good fortune in Korea.)*	12 (4.26)
Metalinguistic information	ChatGPT gives information on language and writing. *(e.g., Subject-Verb Agreement: Ensure that the subject and verb agree in number and tense; Question Hook: Pose a thought-provoking question that encourages readers to consider the impact of technology on children's development.)*	12 (4.26)
Summary	ChatGPT provides a summary of the content of the participant's writing. This usually is a breakdown of the text. *(e.g., Here's the breakdown of the essay: …)*	6 (2.13)
Generation	ChatGPT generates the text upon request, not based on the original text. *(e.g., Certainly, I'll provide responses to two of the findings from the study: …)*	5 (1.77)
Locating	ChatGPT visually highlights the location of what it's indicating with bolded text. *(e.g., Here's the highlighted thesis statement: …)*	1 (0.35)
		Total: 282 (100%)

feedback entirely or followed up by clarifying their previous prompt (also double-coded as 'follow-up' as a prompt type). In cases of 'Modified output,' changes were primarily made to content and vocabulary but less to mechanics, grammar, and structure. 'Full modified output' occurred in 21.5% of the total responses. In other instances, participants chose to type out ChatGPT's revised version rather than copy and paste it. During stimulated recalls, they explained that they wanted to proofread what ChatGPT wrote and selectively replaced some words or portions they didn't want (i.e., selective adjustment). Paraphrasing, or making more extensive modifications (both lexically and syntactically) than in selective adjustment, was observed particularly for B05, who expressed strong motivation to learn English. B05 also mentioned concerns about plagiarism, which further motivated his preference for paraphrasing over directly copying and pasting:

> Researcher: Why did you not copy and paste?
>
> B05: Because I wanted to improve this, like paraphrasing skills. Because it's very important in the research, right? And if you want to copy and paste, I will not improve those skills. (Participant B05)

Table 16.4 *Response types and descriptions*

Response Type	Description	Count (%)
No modified output	The learner does not make any modifications to the text or follow up with another request (double-coded with the prompt type, "Follow-up"	144 (29.27)
Modified output	The learner partially implemented ChatGPT's feedback at five levels:	
	Grammar (e.g., verb tense/form, SV agreement, plurals)	22 (9.09)
	Mechanics (e.g., citation format, punctuation)	33 (13.64)
	Lexical (e.g., replacement of words or phrases)	50 (20.66)
	Structure (e.g., sentence and paragraph structures)	18 (7.44)
	Content (e.g., additions, deletions, modified ideas)	119 (49.17)
	Modified output total	242 (49.19)
Full modified output	The learner fully implemented ChatGPT's feedback (i.e., copy and paste)	106 (21.54)
		Total: 492 (100%)

Table 16.5 *Other response behaviors*

Response Behaviors	Description	Count
Typing	The participant copies and pastes but through typing it in.	16
Viewing	The participant copies and pastes ChatGPT's output above or below their own text during revisions to easily view and implement the feedback into their writing.	12
Reverting	The participant initially does not but later implements the feedback.	11
Comparing	The participant compares ChatGPT's output with their text.	2
		Total: 41

Perceived effectiveness and attitudes

RQ2 addressed the perceived effectiveness of and attitudes towards ChatGPT for language learning. Qualitative insights from semi-structured interviews and stimulated recalls are presented alongside relevant survey response data, with the mean score out of 5 for each variable across three sessions provided in parentheses (e.g., 'local' = x.xx). Note that due to the small sample size, no statistical significance was assumed in the survey results; instead, the survey data serves to complement the qualitative findings by highlighting trends.

Perceived effectiveness

Analysis of semi-structured interviews and stimulated recalls revealed that most participants found ChatGPT's feedback effective at the global level ('global' = 4.17) but less so at the local level ('local' = 3.92) of their writing. Six participants appreciated ChatGPT's effectiveness at the global level. They frequently mentioned its effectiveness in improving structure, overall coherence, and flow of ideas, and particularly addressing the content matter in their writing, such as suggesting elaborations with examples. More importantly, five participants highlighted the benefit of using ChatGPT when integrated into their writing processes for tasks like generating ideas and creating outlines. C07 expressed his struggles with English writing, especially when he had to start with a blank piece of paper. He said ChatGPT helped him overcome that by giving him "a booster" at the beginning stage of writing. B03 and B01 further highlighted ChatGPT's importance in process writing sharing: "I feel ChatGPT is good in the process. [...] It helps you think about things while writing. (Participant B03)." and "For final

editing, I would not use ChatGPT because I feel ChatGPT is a process tool, not an outcome tool. That's why I use it while writing, because it helps me with the process. (Participant B01)."

However, as mentioned by B01 above, participants were not as satisfied with ChatGPT's feedback at the final editing stage and regarding sentence-level errors. Five participants pointed out that ChatGPT was less effective in addressing sentence-level errors, often comparing its performance to other tools such as Grammarly. B01, stating, "grammars were not helpful at all," pointed out that ChatGPT's inability to locate errors might be an issue:

> But I felt like the way that ChatGPT writes and how they edit it, it was like a lot of work to know where the errors were. Whereas with Grammarly, you need to follow each suggestion to see where to compare. (Participant B01)

Four participants expressed mixed feelings about ChatGPT's word choice suggestions, noting issues with repetition, complexity, and inaccuracy. While B01 and C06 found some synonyms and word choices overly complex and repetitive, B03 highlighted inaccuracies in context-specific suggestions, emphasizing a need for more practical and precise word options. Consequently, participants relied on their own judgment in this area, not accepting ChatGPT's suggestions entirely. Instead, they selected appropriate words from the synonym suggestions or replaced overly complex words from ChatGPT's output to make their writing less "artificial." This highlights the limitation of ChatGPT as a chatbot, with participants describing it as "robotic" (B02), a "machine" (C08), and an AI "thing" (B04). As C06 remarked:

> I wish it could diversify words. Because, as a student, the essays should look like I wrote them. But when I use ChatGPT, it kind of shows that ChatGPT wrote it. […] Since it uses certain words repetitively, if the teacher or another person has used ChatGPT before, they can quickly discern that it is a ChatGPT-written essay. (Participant C06)

Most notably, all participants also reported issues with the feedback, such as it being too general and often overly positive, making it less relatable or applicable to their writing. B02 expressed her frustration about general feedback, finding it too lengthy, not understandable, and not pointing out specific problems in her writing. As the feedback got too lengthy and complicated, participants were often observed to skip it entirely: "I was too lazy to read it. I felt like I couldn't really

understand it. I mean, I could, but I felt like I didn't need to. So, I just didn't (Participant C06)."

When the participants only wanted to hear about the weaknesses in their writing to make improvements, ChatGPT often gave them excessive positive feedback, requiring participants to prompt it again for more specific and actionable suggestions. As noted by our participants, this might be because: "[...] ChatGPT wants to keep you positive in general as a human (Participant B01), and thus, "[...] It doesn't tell you. It's not honest. (Participant C07)."

One prominent theme that emerged from the analysis was learning and adaptation in using ChatGPT over time, which may help overcome the limitations in ChatGPT's feedback and thus explain the overall increase in survey scores across different sessions (with the overall 'perceived effectiveness' score improving from 4 to 4.31). Six participants mentioned their realization that the quality of ChatGPT's feedback depended on their prompts and how they improved their prompts to be more specific and straightforward, resulting in more precise and helpful feedback. For instance, B01 noted that he learned to provide smaller text portions instead of the entire text for better results and developed the skill of translating thoughts into clear, actionable prompts. As B04 commented about prompting techniques, "It comes with practice (Participant B04)."

However, some participants reported that ChatGPT's perceived effectiveness did not necessarily lead to actual learning of the language or enhancement in writing skills. Specifically, four participants mentioned "partial" or "no learning" from ChatGPT; while they acknowledged that ChatGPT "helped" with their writing, they could not say they learned from it:

> I would say that ChatGPT gives me an understanding of how to make my paragraphs more descriptive or write more details, but I don't really feel like it helps me to enhance my English. (Participant C07)

Survey responses support this, showing a greater extent of changes based on ChatGPT's feedback ('change' = 4.38), while participants were less able to 'notice' sentence-level errors and global issues that could lead to learning gains ('local' = 3.91; 'global' = 4.17).

Nonetheless, four participants acknowledged that they learned implicitly by comparing their writing with ChatGPT's refined version, as indicated by the following excerpts:

So now I've known how to write this particular sentence in this way, so now I can use it in my future writing. So whenever I have an instance of writing a similar sentence or similar content, I can use the sentences

or words from these versions, and then I can use them for my future writing. (Participant B03)

> Sometimes it might be a small grammatical thing, like I used to write two different words with a hyphen in between. […] I learned that, okay, this is how it is written. And whenever I'm writing now, I'm unconsciously just typing it in the same way ChatGPT does. (Participant B04)

Attitudes towards ChatGPT

All participants had at least some trust in ChatGPT and its capabilities ('trust' = 3.46). Those with strong trust, like B04 and C06, highlighted its accuracy, effectiveness with specific prompts, and overall reliability for certain tasks like English writing (as opposed to, for instance, math problems). For instance, B04, identified as an experienced and skilled user in our observation of screen recordings, expressed the highest level of trust among all participants. When asked whether there were any instances of feedback that were not helpful, he said: "It was all the prompts I used that I knew would work perfectly there, so I tried to use them instead of other prompts (Participant B04)," suggesting that perceived effectiveness and attitudes can improve with increased proficiency in prompting skills. B04 seemed to have strong confidence in his prompting skills and awareness of common system errors and bugs in ChatGPT so that even in cases where he encountered such problems, he could make things work out or avoid writing the prompts that might cause those issues in the first place.

However, two participants expressed reduced trust in ChatGPT due to inconsistent, irrelevant, and inaccurate responses, as shared in these two excerpts:

> And if I ask a question and then I know that question is not correct, if I again ask ChatGPT, 'Are you sure?' then I see that it changes the answer and says, 'I'm sorry for that sentence.' So, I don't completely trust ChatGPT. (Participant C07)
>
> When I put things in ChatGPT, it tries to take it away from what I'm trying to write and posts a bunch of stuff from the internet, which I don't think fits in there. […] I can't really trust everything because I have to look it up and manage it that way. (Participant C08)

As indicated by C08 in the previous excerpt, participants said they tended to double-check with other sources or tools as a cautious approach. Even B04, who expressed very strong trust in ChatGPT,

mentioned that he used Grammarly for final proofreading even after ChatGPT.

When asked about their preference between teacher feedback and ChatGPT's feedback, five participants preferred teacher feedback ('preference' for ChatGPT = 3.25), finding it more specific and contextually informed. C08, referring to ChatGPT as a "machine" said "Humans, they know better. I still say they know better than machines, they will be able to make you understand the things they are saying. (Participant C08)." However, even those who favored teacher feedback acknowledged ChatGPT's advantage of increased consistency and accessibility, allowing them to ask for feedback as many times as they wanted. B01 noted:

> But also, it depends on the teacher. Some teachers will not be. So, the good thing about ChatGPT is that it's consistent. Even if it's not the best quality, it's always there. And that's something that gives it a big advantage. (Participant B01)

Despite its limitations, all participants expressed a desire to continue using ChatGPT ('future' = 3.83), whether strongly or, to some extent, viewing it as a valuable supplementary tool in their writing process.

Discussion

This study examined how L2 English learners interacted with ChatGPT in terms of prompt, feedback, and response types across three revision sessions, as well as their perceived effectiveness of ChatGPT as an AWE tool. In answering RQ1, we found that 'Revision request' was the second most used prompt type, often leading to full modified output as observed in the screen recordings. However, contrary to the positive implication of the term, 'full' was not ideal, as it indicated that participants had directly copied and pasted ChatGPT's revised text without making any modifications on their end. Interestingly, in many cases, even when students asked for feedback, ChatGPT provided fully revised versions of their drafts. As a result, many participants skipped reading the feedback and simply adopted the revised version. However, a full or near copy of ChatGPT's output can hardly be considered modified output (Keck, 2014), thereby limiting learning opportunities. This also raises plagiarism concerns caused by ChatGPT, as noted by the students in our study and in previous research (Xiao & Zhi, 2023). Nevertheless, about half of the responses (49.19%) resulted in modified output, where students implemented the feedback on their own, particularly with content-level changes (49.17% across all five levels; grammar, mechanics, lexical, structure, and content). This suggests

the potential for learners' engagement with ChatGPT to contribute to writing skills development. As some participants noted in interviews, this might not result in explicit language learning; however, implicit learning may still occur through repeated exposure to correct language forms in ChatGPT's feedback.

In answering RQ2, we found participants often found ChatGPT's feedback too general and overly positive, making it less actionable for students. Additionally, the feedback was often lengthy and complicated, which may have increased the cognitive load on the learners and complicated comprehension (Ranalli et al., 2017; Weigle, 2013). This may have further reduced ChatGPT's LLP, leading participants to take shortcuts rather than engage with the feedback. However, as participants mentioned in interviews, the effectiveness of feedback would depend on how much the learner cares about the text and their motivation to learn the language. More motivated learners are more likely to attend to ChatGPT's feedback (Xiao & Zhi, 2023).

Regarding local versus global aspects of writing, ChatGPT was perceived as less effective in addressing local errors due to its limited perfection, which does not align with previous research (Ali et al., 2023; Mahapatra, 2024; Song & Song, 2023). This perception is supported by findings from RQ1, which showed relatively low percentages of modified output at the grammar (9.09%) and mechanics (13.64%) levels. Unlike tools such as Grammarly, which provides visual aids like flagging and highlighting, ChatGPT lacks features that make it easier for learners to identify where errors occurred, or changes were made. Integrating these types of visual aids with ChatGPT's feedback could help learners produce more appropriately modified outputs, leading to more effective learning outcomes.

In terms of global aspects, learners found ChatGPT effective in refining their drafts by improving the flow of ideas and enhancing overall delivery; this aligns with previous literature (Boudouaia et al., 2024; Mahapatra, 2024; Song & Song, 2023). Unlike previous AWE tools, which excelled at addressing sentence-level correctness, ChatGPT's more advanced technology allows it to provide customized feedback on global writing concerns, adapting to each student's unique text. Therefore, integrating ChatGPT into classroom teaching could further reduce teachers' workloads while enhancing students' L2 writing development by providing continuous access to an AI writing assistant.

Moreover, despite not being directly related to LLP, our findings suggest the potential for using ChatGPT to promote a process-oriented writing approach in classroom settings. As highlighted by our participants, students can use ChatGPT at any stage of their writing process as a personal writing assistant. They found it particularly useful when

beginning their writing, as it helps in generating ideas on the topic and organizing thoughts by outlining their drafts.

Given these benefits, despite ChatGPT's limitations, we suggest that teachers collaborate with ChatGPT in providing feedback on student writing. Teachers should aim to help students set up ChatGPT to become an effective tool that complements their feedback and maximizes actual learning. We acknowledge that the use of AI in student work poses challenges to academic integrity, as evidenced in our study where students often copied and pasted directly from ChatGPT into their writing. Therefore, it is the responsibility of instructors to decide whether to allow the use of ChatGPT in their courses. However, as it is nearly impossible to ban students from using ChatGPT without a reliable means of detecting LLM-generated content, the key question concerns how teachers can guide students to use such tools in a legitimate way. Based on our findings, we encourage teachers to consider the following points to guide students:

- Encourage students to practice their prompting skills to make their prompts more specific and straightforward. Emphasize that it is really a communication between them and ChatGPT, and the results depend on their prompt. For example, instead of prompting, "Give me feedback on this writing," more specific prompts indicating which aspects they want feedback on work better, such as "Give me feedback on the flow of this writing." Students can also provide context or the purpose of their writing to get more individualized feedback. If their prompt does not work out, students should be encouraged to negotiate the meaning with ChatGPT through iterative interactions until they are satisfied with the results.
- Teachers should emphasize using ChatGPT as a learning tool to assist students throughout the entire writing process. Students can generate ideas and outlines to ease the burden of writing from scratch. During the writing stage, students should be guided to ask for feedback on various aspects of their writing, from mechanics and grammar errors to more global aspects, including connectivity between sentences and overall coherence. Let students be cautious when using ChatGPT for final proofreading, however, as sometimes, ChatGPT was found to have errors. They should double-check with other tools that are more specialized for those purposes.
- To increase the noticing of errors for actual learning gains, students can ask ChatGPT to locate the errors or changes made using a simple prompt like "highlight the errors/changes."
- Teachers should discuss with students when and when not to use ChatGPT for their writing. Students should be cautioned against

copying and pasting behavior. Rather than generating the full draft or seeking a fully revised text from ChatGPT, students should view ChatGPT as a personal writing teacher that is easily accessible as needed for their writing process. Teachers should make it explicit that misuse of ChatGPT, such as copying and pasting, constitutes cheating and plagiarism, as it involves presenting ChatGPT's output as their own work, and there would be a lack of learning from it.

- Teachers can provide students with default prompts to be used with the customization feature in ChatGPT, which would yield more structured feedback. This feature can be found under Profile > Customize ChatGPT. Refer to Figure 2 below for an example:

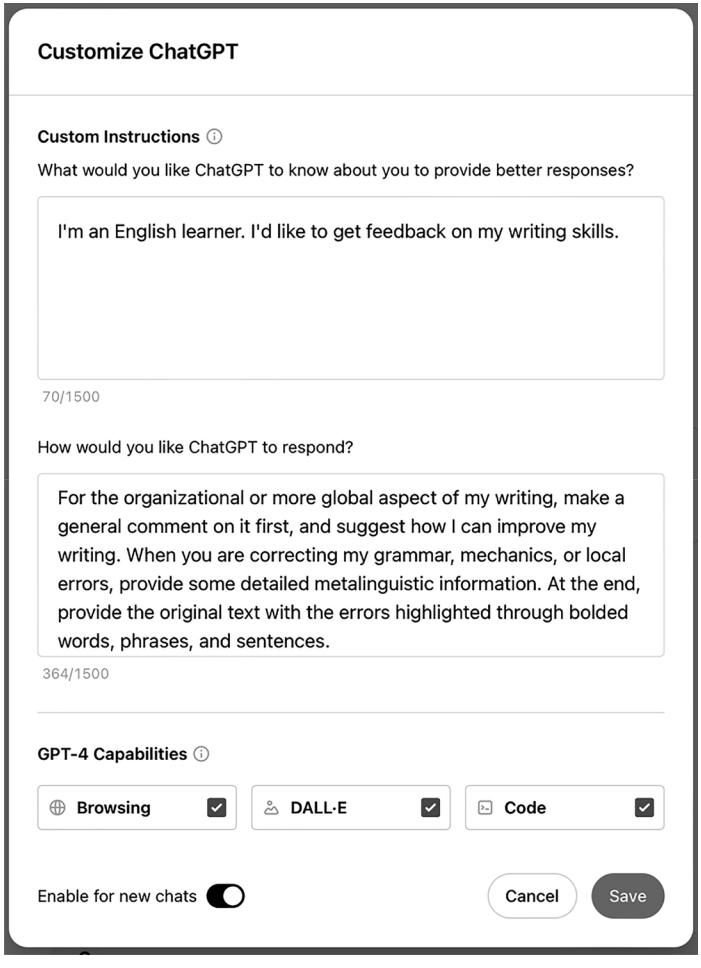

Figure 16.2 *Customization feature in ChatGPT.*

Noting that this study has some limitations is important for future research. First, our sample size is very small, with only eight participants, which limits the generalizability of our findings. More participants are needed to improve this aspect. Second, in our effort to explore overall trends, we did not fully focus on individual participants when reporting our findings. Future research should include a more detailed qualitative exploration of each individual learner as a case study. Last, our study was conducted over a short period, which may not have been sufficient to observe changes in learners' interaction patterns and perceptions of ChatGPT. Future research should include longitudinal studies to address this gap. Despite these limitations, our exploratory study aims to provide a foundation for researching the most recent AI technology as an AWE tool. Additionally, it aims to help teachers guide students to use ChatGPT effectively, enhancing their language learning potential.

Conclusion

This chapter explores the language learning potential of ChatGPT, focusing on how L2 students interacted with ChatGPT (prompt type, feedback type, and response type) and their perceptions of its effectiveness and attitudes towards using ChatGPT for English writing. Findings indicate that the learning potential of ChatGPT is limited, as students often asked ChatGPT to produce revised drafts, frequently copying, pasting, and merely replacing some words from the generated text. Feedback was often found overwhelming or predominantly positive, making it less applicable for revising their writing. While less effective in addressing local errors, L2 students appreciated the value of ChatGPT for more global concerns of writing and a process-oriented writing approach, particularly for generating ideas and outlining at the beginning stage of their writing. We hope that the findings of our study suggest pedagogical implications for instructors intending to use ChatGPT in language classrooms, facilitating its effective implementation.

Post-reading questions:

1) Based on the findings, how do learners' interactions with ChatGPT vary in terms of prompt types and feedback received?
2) What were some common issues learners experienced with ChatGPT's feedback, and how did they address them?
3) What are the key recommendations for teachers to help students use ChatGPT more effectively in their writing processes?

Further reading

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

Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies*, *28*(11), 13943–13967. https://doi.org/10.1007/s10639-023-11742-4

References

Ali, J. K., Shamsan, M. A., Hezam, T. A., & Mohammed, A. A. (2023). Impact of ChatGPT on learning motivation. *Journal of English Studies in Arabia Felix, 2*(1), 41–49. https://doi.org/ 10.56540/jesaf.v2i1.51

Boudouaia, A., Mouas, S., & Kouider, B. (2024). A study on ChatGPT-4 as an innovative approach to enhancing English as a foreign language writing learning. *Journal of Educational Computing Research, 62*(6), 1509–1537. https://doi.org/10.1177/07356331241247465

Chapelle, C. A. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing, and research*. Cambridge University Press.

Chapelle, C. A. (2007). Technology and second language acquisition. *Annual Review of Applied Linguistics, 27*, 98–114. https://doi.org/10.1017/S0267190508070050

Chen, Z., Chen, W., Jia, J., & Le, H. (2022). Exploring AWE-supported writing process: An activity theory perspective. *Language Learning & Technology, 26*(2), 129–148. https://doi.org/10125/73482

Corbin, J., & Strauss, A. (2007). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (3rd edition). Sage.

Cotos, E. (2023). Automated feedback on writing. In *Digital writing technologies in higher education: Theory, research, and practice* (pp. 347–364). Springer International Publishing.

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), 57. https://doi.org/10.1186/s41239-023-00425-2

Fuchs, K. (2023). Exploring the opportunities and challenges of NLP models in higher education: Is ChatGPT a blessing or a curse? *Frontiers in Education*, *8*, 1166682. https://doi.org/10.3389/feduc.2023.1166682

Gass, S., & Mackey, A. (2006). Input, interaction and output: An overview. In K. Bardovi-Harlig and Z. Dörnyei (Eds.), *Themes in SLA research* (pp. 3–17). John Benjamins.

Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology, 2*(26), 5–24. https://doi.org/10125/73474

Guo, K., & Wang, D. (2023). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*,1–29. https://doi.org/10.1007/s10639-023-12146-0

Keck, C. (2014). Copying, paraphrasing, and academic writing development: A re-examination of L1 and L2 summarization practices. *Journal of Second Language Writing*, *25*, 4–22. https://doi.org/10.1016/j.jslw.2014.05.005

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

Long, M. H. (2007). *Problems in SLA*. Lawrence Erlbaum.

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments, 11*(1), Article 9. https://doi.org/10.1186/s40561-024-00295-9

Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology*, *37*(1), 8–25. https://doi.org/10.1080/01443410.2015.1136407

Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing, 52*, 100816. https://doi.org/10.1016/j.jslw.2021.100816

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

Stevenson, M., & Phakiti, A. (2019). Automated feedback and second language writing. In K. Hyland, & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 125–142). Cambridge University Press.

VERBI Software. (2024). MAXQDA 2024 [computer software]. VERBI Software. Available from maxqda.com

Wang, Z. (2022). Computer-assisted EFL writing and evaluations based on artificial intelligence: A case from a college reading and writing course. *Library Hi Tech, 40*(1), 80–97. https://doi.org/10.1108/LHT-05-2020-0113

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal*, *3*(1), 22–36.

Weigle, S. C. (2013). English as a second language writing and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluations: Current applications and new directions* (pp. 36–54). Routledge.

Xiao, Y., & Zhi, Y. (2023). An exploratory study of EFL learners' use of ChatGPT for language learning tasks: Experience and perceptions. *Languages, 8*(3), 212.

Zhang, Z., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing*, *36*, 90–102. https://doi.org/10.1016/j.asw.2018.02.004

Zhang, Z. (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing, 43*, 100439 https://doi.org/10.1016/j.asw.2019.100439