





ARTICLE



<https://doi.org/10.1057/s41599-024-03543-y>

OPEN

# L2 writer engagement with automated written corrective feedback provided by ChatGPT: A mixed-method multiple case study

Da Yan <sup>1</sup> & Shuxian Zhang<sup>1</sup> 

Automated written corrective feedback (AWCF) has been widely applied in second language (L2) writing classrooms in the past few decades. Recently, the introduction of tools based on generative artificial intelligence (GAI) such as ChatGPT heralds groundbreaking changes in the conceptualization and practice of AWCF in L2 pedagogy. However, students' engagement in such an interactive and intelligent learning environment remains unstudied. The present study aims to investigate L2 writers' behavioral, cognitive, and affective engagement with ChatGPT as an AWCF provider for writing products. This mixed-method multiple case study explored four L2 writers' behavioral, cognitive, and affective engagement with AWCF provided by ChatGPT. Bearing the conversational and generative mechanisms of ChatGPT in mind, data on students' engagement were collected from various sources: prompt writing techniques, revision operations, utilization of metacognitive and cognitive strategies, and attitudinal responses to the feedback. The results indicated that: 1) behavioral engagement was related to their individual differences in language proficiencies and technological competencies; 2) the participants have failed to metacognitively regulate the learning processes in an effective manner; and 3) ChatGPT ushered in an affectively engaging, albeit competence-demanding and time-consuming, learning environment for L2 writers. The study delivers conceptual and pedagogical implications for educators and researchers poised to incorporate GAI-based technologies in language education.

<sup>1</sup>School of Foreign Languages, Xinyang Agricultural and Forestry University, Xinyang, China. email: [shuxianzhang@aliyun.com](mailto:shuxianzhang@aliyun.com)

## Introduction

“Engagement defines all learning” (Hiver et al. 2021, 2).

In second language (L2) writing, feedback, especially written feedback, is one of the most widely applied and researched topics (Hyland and Hyland 2019). In the past decades, the focus of relevant research has shifted from the effects of feedback on writing quality (e.g., Nelson and Schunn 2009; Dizon and Gayed 2021) towards students’ involvement in processing and utilizing feedback (Zhang 2017; Ranalli 2021). However, due to the multifaceted and dynamic nature of student engagement with written feedback (Han and Gao 2021), the body of existing literature suffers from the lack of multidimensional insights into all the aspects of engagement with feedback (Shi 2021).

Meanwhile, with the advancement of technologies, automated written corrective feedback (AWCF) has been widely implemented in L2 classrooms as a pedagogical innovation. Researchers have made continuous contributions to expand our knowledge in 1) the effects of AWCF on the quality of writing products (Barrot 2021); 2) the interplay of AWCF and classroom instruction (Tan et al. 2022); and 3) learners’ perceptions of the utilization of AWCF providers in L2 classrooms (O’Neill and Russell 2019). Reversely, thorough investigations of students’ engagement with AWCF have been scant (Koltovskaia 2020). Furthermore, compared to the bravery to incorporate state-of-the-art technologies in L2 classrooms, there remains a lacuna of research on the students’ engagement with cutting-edge AWCF providers. Since its advent in late 2022, ChatGPT, a conversational generative artificial intelligence (GAI) chatbot powered by large language models (LLM), has evoked heated hype about its impact on language education (e.g., Jiao et al. 2023; Mizumoto and Eguchi 2023). Specifically, a few pioneering studies have unveiled its strength to outperform its precedents in correcting grammatical errors (Fang et al. 2023; Wu et al. 2023). Nevertheless, we confronted a dearth of empirical evidence of students’ engagement with AWCF generated by ChatGPT in authentic L2 pedagogical settings.

Against the above backdrops, the study has explored L2 writers’ engagement with AWCF provided by ChatGPT. Theoretically, the research has drawn upon existing studies to reconceptualize student engagement with feedback provided by GAI-based systems. Methodologically, the study adopted a mixed-method multiple case study approach to collect and triangulate data. The paper is significant as it brings new insights into the changes in learning patterns that resulted from students’ exposure to GAI-based feedback providers and the extent to which learners engage with the new environment.

## Literature review

**AWCF and the potential of ChatGPT.** In recent years, the impact of AWCF, the written corrective feedback (WCF) provided by computerized automated writing evaluation (AWE) tools, on L2 writing pedagogy has grown continuously (Zheng et al. 2021). Compared to the traditional teacher-fronted WCF, AWCF has been praised by researchers and educators for its: 1) power to alleviate teachers’ and peers’ burden in L2 classrooms (Ranalli 2018); 2) empowering effects in augmenting students’ involvement in revision and proofreading (Li et al. 2015); and 3) promptness in providing effective feedback (Barrot 2021). However, researchers have conflicting perspectives regarding the efficacy of AWCF compared to WCF. On the one hand, technology-enhanced feedback providers or interventions serve as a significant assistant in facilitating teachers or peers in making an accurate evaluative judgment on writing artifacts, particularly in overcoming evaluation biases or inaccuracies (Wood 2022; Gong and Yan 2023; Yan 2024a), for example, the choice between

lenient or severe judgment (e.g., Jansen et al. 2021) or the tendency to use simple heuristics while forming feedback (e.g., Fleckenstein et al. 2018). On the other hand, AWCF has constantly been criticized as inferior to human-generated feedback with the relatively restricted abilities of AWE systems to form accurate and comprehensive evaluations of writing artifacts, particularly the more traditional corpus-based systems such as Pigai.com (Fu et al. 2022). Hence, there has been a long-standing pursuit to improve AWE systems in providing individualized and effective AWCF for language learners (Fleckenstein et al. 2023).

Recently, with the emergence of AI-based technologies such as Grammarly and QuillBot, researchers’ interest shifted gradually. According to existing empirical studies, AI-based AWCF providers outperform the corpus-based systems by a substantial margin in both the feedback uptake and revision quality of L2 writers (c.f., the successful revision rate of merely 60% in Bai and Hu 2017; and approximately 70% in Koltovskaia 2020). Based on such improvement in performance, the technological advancement would further spur the research and implementation of AWCF providers in L2 writing classrooms.

Since the appearance of ChatGPT, researchers have attempted to adopt it as an AWCF provider for L2 writing with promising results. As evidenced by the comparison between ChatGPT and Grammarly by Wu et al. (2023), the former offers a further improvement over existing AI-based solutions for correcting grammatical errors. Accordingly, researchers have optimistically prophesied the potential of ChatGPT as a significant assistant for language learners in the future (Jiao et al. 2023; Mizumoto and Eguchi, 2023). The potential of ChatGPT as a potential AWCF provider is based on: 1) the outstanding performance in providing grammatical and syntactical corrections in an accurate and instant fashion (Steiss et al. 2024); 2) the tremendous amount of pre-trained language data that ensures its excellent performance compared to its precedents (Wu et al. 2023); 3) the ability to iteratively respond to users’ inquiries for feedback due to the interactional and conversational mechanism of the human-computer interface (White et al. 2023; Yan 2024b); and 4) the verified enhancement from conversational AI-based chatbots as learning assistants in previous studies (Wu and Yu 2023).

However, we cannot neglect that ChatGPT has its disadvantages; for example, it could create *hallucination*, the randomly generated and unverified information (Tonmoy et al. 2024). Additionally, since ChatGPT is a conversational chatbot, the quality of ChatGPT-generated feedback is dynamic and subject to the extent to which the learners *agentically* seek and process the feedback (Yan 2024b). Moreover, from a student perspective, the effective and ethical use of ChatGPT called for a higher level of AI literacy and corresponding support and scaffolding from teachers or peers, both of which were inadequately possessed or provided at the current stage (Yan 2023). Taken together, the effective utilization of ChatGPT in educational settings needs meaningful and successful fulfillment of its potential while controlling the threats and menaces it might bring.

In the pre-ChatGPT era, Ranalli (2018) has called for an accurate and robust AWCF provider that could interactively answer individual learners’ specific needs and demands. Given the history of the AWCF application and the strength of ChatGPT, the GAI-based system is in the spotlight as a potential problem solver and game changer for the field.

In an era of change, the effects of ChatGPT or similar GAI-based tools on L2 writing still need to be studied. Among all the overheating hype and unfounded fears about adopting ChatGPT in education since its debut, we expect more empirical studies investigating the actual effects of the tool on language learners. As Zhang (2017) has suggested, students’ engagement with feedback

providers is an indispensable prerequisite to benefiting from technology-mediated language learning facilities. Consequently, a study focusing on learners' involvement in processing and utilizing the corrective feedback provided by ChatGPT would enrich our limited knowledge of AI-mediated language learning (e.g., Tseng and Warschauer 2023).

**Student engagement with AWCF and relevant empirical research.** In L2 research, engagement has been understood as one of the defining features of students' active involvement in learning (Mercer 2019). For L2 writing, engagement is commonly conceptualized as a tripartite meta-construct composing three key components: behavioral, affective, and cognitive engagement (Ellis 2010; Zhang and Hyland 2018; Fan and Xu 2020). Specifically, behavioral engagement refers to the learning behaviors (Zheng and Yu 2018) and strategic choices in translating the received feedback into a revision (Han and Hyland 2015); affective engagement represents students' emotional and attitudinal responses to the feedback (Ellis 2010); and cognitive engagement denotes the extent to which the student cognitively perceives the feedback and the subsequent cognitive and metacognitive operations to process and utilize the feedback (Han and Hyland 2015).

In recent years, many studies have investigated the three dimensions of student engagement in pedagogical settings of L2 writing equipped with automated feedback providers. On the one hand, researchers have attributed students' engagement with AWCF to various factors. In a single case study to examine engagement with Pigai.com in an EFL context, Zhang (2017) discovers that more teacher scaffolding and pedagogical assistance are needed to facilitate the cognitive engagement of L2 writers learning with AWE systems. In a subsequent multiple case study on engagement with teacher-scaffolded feedback provided by Pigai.com, Zhang and Hyland (2018) attribute the diversity in learners' engagement to students' language proficiency, learning styles, and utilization of learning strategies. As the interest of researchers shifts from traditional AWE systems to AI-based AWCF providers, new perspectives on student engagement emerge. Ranalli (2021) concludes by observing six Mandarin L1 learners who trust in AWCF quality and credibility and decisively determine engagement. Furthermore, a recent eye-tracking study reveals that feedback explicitness determines student engagement with AWCF provided by Write & Improve (Liu and Yu 2022). On the other hand, contradictory voices are often heard from research on the students' engagement with AWCF. For example, the study by Rad et al. (2023) betokens the promoting effects of Wordtune, an AI-based writing assistant, on L2 students' overall engagement. On the contrary, Koltovskaia (2020) manifests that students' cognitive engagement with the feedback provided by Grammarly is insufficient, although positive affective engagement was reported after using the tool to support writing.

Despite the prolific insights into students' engagement with AWCF in L2 writing classrooms, scholars have criticized existing research for neglecting key elements, e.g., overlooking students' involvement in the revision process (Stevenson and Phakiti 2019), and the predominance of an outcome-based approach to studying the quality of writing products (Liu and Yu 2022). The present study not only embarks on a comprehensive investigation into students' engagement but also strives to seek a new conceptual departure in L2 pedagogy in the age of AI. Considering the characteristics of ChatGPT as a potential AWCF provider, there exists a lacuna in our understanding of how and to what extent students engage with the new GAI-based feedback provider.

### Conceptualizing engagement with GAI-generated feedback.

The rationale to revisit the conceptualization of student engagement with corrective feedback in the context of GAI is posited on the paradox between the alleged positive effects of AWCF providers on writing pedagogy (Fang et al. 2023; Wu and Yu 2023) and the reported challenges encountered by students to effectively tap the strength of AI in seeking feedback (Yan 2024b). To frame the decisive factors affecting engagement, Ellis's (2010) componential framework for investigating corrective feedback is referred to. According to the framework, student engagement with corrective feedback is influenced by individual differences and contextual factors. Previous studies have generally attributed the individual differences of learners to language proficiency (Zhang and Hyland 2018; Ranalli 2021). However, for ChatGPT as an AWCF provider, technological competence should be included as a major aspect of individual competence since the interaction with ChatGPT, via iteratively prompt writing and amendments, calls for a higher level of digital literacy (Lee 2023; Naamati-Schneider and Alt 2024).

The tripartite dimensions within the meta-construct of engagement are developed on top of the body of literature. First, the concept of behavioral engagement is expanded. In the study of Zhang and Hyland (2018), behavioral engagement is deemed to be students' behaviors to process feedback, i.e., operation and strategies of revision. However, for the present study, an additional aspect of students' behaviors is considered, i.e., the actions of writing prompts to seek feedback from ChatGPT. Unlike conventional AWE systems and AWCF providers such as Grammarly, the quality, content, and quantity of feedback provided by ChatGPT rely on the user's interaction with the GAI-based system through iterative and incremental prompt writing (Yan 2023). Second, in line with the work by Koltovskaia (2020), the present study conceptualizes cognitive engagement as students' utilization of cognitive and metacognitive strategies in processing AWCF; and affective engagement as students' emotional and attitudinal responses to the AWCF. The conceptual model of student engagement with ChatGPT-generated feedback is graphically shown in Fig. 1.

**The current study.** The study explores L2 writer engagement with AWCF generated by ChatGPT. The following logic guides the research: (1) compared to more traditional approaches to corrective feedback, we are facing a paucity of comprehensive understanding of student engagement with AWCF; (2) compared to AWE systems such as Pigai.com, we have barely any knowledge about how ChatGPT's unique features, such as its outstanding text generation abilities, interactive and conversational interfaces, iterative feedback generation capabilities, would impact on L2 writer engagement with AWCF; and (3) given that effective use of ChatGPT calls for a higher level of domain knowledge and AI competence, we need to examine how do these individual characteristics influence L2 writer engagement with ChatGPT-generated AWCF. Therefore, the following research question would be answered:

*How do L2 writer with varied language proficiency and technological competence behaviorally, cognitively, and affectively engage with AWCF provided by ChatGPT?*

### Methods

**Participants.** The study's research site was an undergraduate EFL program at a Chinese university. Students enrolled in this program had to take three writing courses in which formative assessment and technology-enhanced feedback were practiced. Therefore, the students were relatively experienced in learning-oriented assessment practices.

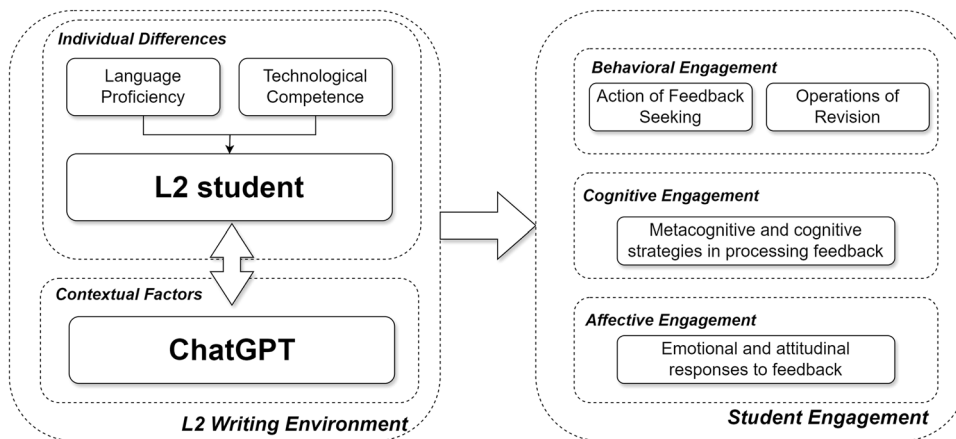


Fig. 1 Conceptual model of student engagement with ChatGPT-generated AWCF.

Table 1 Background information of participants.

Pseudonym	Gender	Age	Years of learning L2	Language proficiency <sup>a</sup>	Technological competence <sup>a</sup>	Recommendations from co-researchers
Emma	Female	19	10	92/100	86/100	"A must choice for participants in the study."
Sophia	Female	21	12	87/100	71/100	"Typical high-proficiency language learners."
Robert	Male	20	11	72/100	88/100	"A representative for male students of language majors."
Mia	Female	20	11	74/100	71/100	"A typical hard-working yet medium-proficiency student."

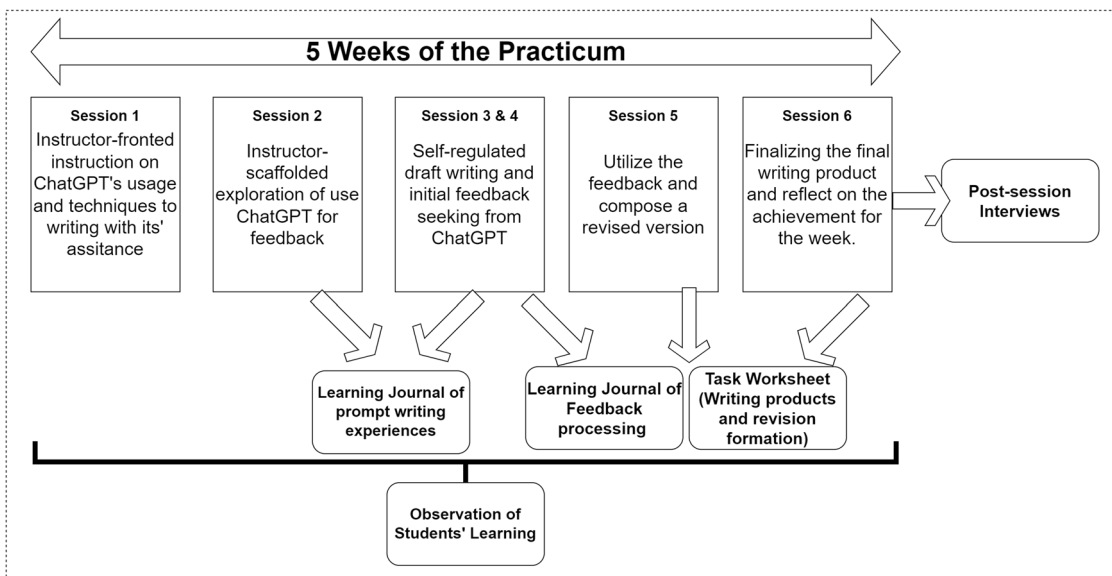
<sup>a</sup>Language proficiency: mean score of four precedent L2 writing assessments;  
Technological competence: mean score of precedent assessments in two digital humanities courses.

The participants were recruited from a pool of students previously involved in a pilot project investigating the impact of ChatGPT on L2 learners (Yan 2023). A purposeful sampling method was applied to select four participants with distinct characteristics in language proficiency and technological competence (Palinkas et al. 2015). The sampling criteria included: 1) average performance in four precedent L2 writing assessments, which were adopted from the official writing prompts of Test for English Major band 4, a national level and widely applied test of English proficiency for English majors in China (Jin and Fan 2011); 2) average performances in the assessments of two precedent digital humanities courses; 3) interest in the project and self-rated trust in AWCF; and 4) recommendations from co-researchers (from the teaching faculty of the program) based on classroom observation and the analysis of learning artifacts. Originally, a group of 14 students voluntarily participated in the project. However, only 4 students were regarded as qualified participants for the present study since the others failed to provide complete learning data. See Table 1 for the background information of the 4 participants. To maintain the ethicality of the study, written informed consents were obtained from all participants, who were aware of the purpose, design, procedures, and anonymity policies of the study, prior to the data collection procedures.

**Design.** In second language acquisition (Duff 2010) and educational feedback (e.g., Zhang and Hyland 2023), case study has been widely applied as an established means to collect rich data on students' actual learning experiences. Adopting a mixed-method multiple case study approach (Yin 2013), a case in the study was defined as the extent to which an individual learner was behaviorally, cognitively, and affectively engaged with ChatGPT-

generated feedback. For each specific case, the study followed a convergent design in which the quantitative and qualitative data were triangulated to manifest students' engagement with the AWCF (Creswell and Plano Clark 2018). Furthermore, the study was a collective multiple-case study, as the cross-case comparison of the individual cases allowed the researcher to generalize the findings for a broader context (Stake 1995). Although the limited number of participants would possibly hinder the study's potential implications for a general and broader context, small sample size and/or high drop-out rate are frequent phenomena among case studies on learning behaviors, for example, in Koltovskaia and Mahapatra (2022), only 2 student participants' data were selected from a pool of 17; in Yan (2024b), only 3 students were finalized as participants in the inquiry into L2 writer's feedback-seeking behaviors. As argued by Adams (2019), the limited number of research subjects in case studies had its merits in unfolding learner experiences in using feedback other than the feedback design.

**Procedures.** During the five-week project, 68 students (inclusive of all the participants of the study) joined an L2 writing practicum focusing on exploring the affordance of ChatGPT as a feedback provider. Each week, two sessions of teacher-fronted instruction and live demonstration were prescribed, in addition to four sessions of self-directed learning and practicing. Students must complete draft writing, seek feedback from ChatGPT, execute the revision based on the feedback, and submit it to the instructor each week. To facilitate the data collection, different data collection strategies were employed, i.e., students' weekly reflective learning journals (Bowen 2009), the observation of students' behaviors in the classroom (Jamshed 2014), and the interviews (Braun and Clarke 2012). The practicum structure and data collection procedures of the study are shown in Fig. 2.



**Fig. 2** Procedures of the study.

First, after each week, the participants were required to complete a reflective learning journal. Specifically, they are asked to provide their weekly reflection on the learning progressions, experiences using ChatGPT for feedback, the episodes of interaction with ChatGPT for eliciting and refining corrective feedback, and the acceptance and rejection of the feedback in preparing the revisions. Participants were encouraged to complete the journal multimodally with multiple types of files, e.g., screenshots, audio recordings, and video clips as supplementary files. See Supplementary Appendix A for the template for the reflective journal. Moreover, a task worksheet was provided to the learners to write down the draft writing, formative revisions, and the final writing products for each writing task. See Supplementary Appendix B for a sample task worksheet.

Second, during each instructional and practice session, the instructors were requested to record the students' learning behaviors and processes. The students attended all the sessions in language laboratories equipped with keylogging and screen recording facilities to facilitate the recording. All the loggings and recordings were gathered, processed, and taken down in notes by two co-researchers recruited from the teaching faculty. Furthermore, the note-takers coded the notes against a coding scheme for metacognitive and cognitive learning strategies for the study. See Supplementary Appendix C for the coding scheme adopted from the work of Sonnenberg and Bannert (2015). Inter-coder disagreements were solved by reaching a consensus between the two coders and the researcher through recording playbacks and collective discussion. According to the measurement of Cohen's Kappa ( $\kappa = 0.72$ , 95% CI [0.65, 0.84]), good inter-rater reliability was attained.

Finally, an immediate post-session interview was performed for each participant after the final session of the week. Participants were required to follow the instructions of the interviewer to answer the questions from the pre-determined interview protocol with questions like "When the project ends, are you willing to continue using ChatGPT for feedback in L2 writing?". Each interview session lasted for about 10–15 min. The moderator was required to write down all the major viewpoints and interview details in an interview note. The interviews were audio recorded and transcribed verbatim.

Throughout the data collection processes, the researchers have taken measures to ensure the trustworthiness, reliability, and

validity of data. For example, the reliability and validity of the observational data were attained after reaching interobserver agreement during the initial two weeks (Watkins and Pacheco 2000); the trustworthiness of the qualitative data were checked with member checking with participants (Doyle 2007), and the investigator triangulation (Carter et al. 2014).

**Analysis.** First, quantified document analysis was applied to analyze the learning journals and the worksheets. For each case, the individual learner's learning details, i.e., time spent for feedback processing, number of written ChatGPT prompts, time spent for the interaction with ChatGPT, and retention of feedback in the revision, were quantified and analyzed through descriptive statistics. For the coding of prompt writing patterns, a coding scheme developed by the first author in a previous work was used (Yan 2024b). The coding was performed by three coders recruited from the teaching faculty. According to the measurement of Fleiss' Kappa ( $\kappa = 0.86$ , 95%CI [0.78, 0.91]), good inter-rater reliability was achieved.

Second, a lag sequential analysis (LSA) using GSEQ 5.1 software was performed to analyze students' transition and interaction patterns using metacognitive and cognitive strategies extracted from the coded classroom observations. LSA is a statistical technique used to identify patterns and sequences of behaviors or events over time by examining the conditional probabilities of one event occurring after another within a specified time delay or lag period (Bakeman and Quera 2011). Correspondingly, GSEQ calculated adjusted residuals from a transitional probability matrix based on the coded behavior sequences (Pohl et al. 2016). The significance of behavioral transitions was determined by the Z-score of the adjusted residuals (significant if  $Z > 1.96$ ). Behavioral transitions were visualized to present the behavioral patterns in terms of metacognition and cognition within the feedback processing and revision processes.

Third, a thematic analysis following the six-step procedures recommended by Braun and Clarke (2012) was applied to the interview transcripts. Two additional co-researchers were recruited to assist the researchers in coding and theme extraction. Disagreements among the co-researchers were solved through an ad hoc discussion convened and joined by the researcher.

Finally, when the data analyses were finalized, all findings were converged and triangulated to answer the research questions.

**Results**

**Behavioral engagement.** After the quantified document analysis, the data on the participants’ feedback-seeking and revision operations were presented. Specifically, the actions of feedback seeking and revision were respectively manifested as detailed patterns in composing ChatGPT prompts and processing ChatGPT-generated feedback categorized by error types.

First, the actions of feedback seeking by the four participants were shown in Figs. 3–6, respectively. According to the bar charts, Emma and Sophia created more than 2000 ChatGPT prompts in 5 weeks, followed by Robert’s 1670 and Mia’s 1238. Pertinent to the weekly developmental trends in using specific prompt writing techniques, Emma and Sophia have displayed similarities, indicating that the patterns of Robert and Mia were on common ground. For example, in using the [+QUA] technique (providing the user’s quality evaluation of the feedback to re-elicite feedback from ChatGPT), Emma and Sophia have displayed a parabolic curve in the weekly frequencies. At the same time, Robert and Mia have kept a growing momentum to use such a technique throughout the project.

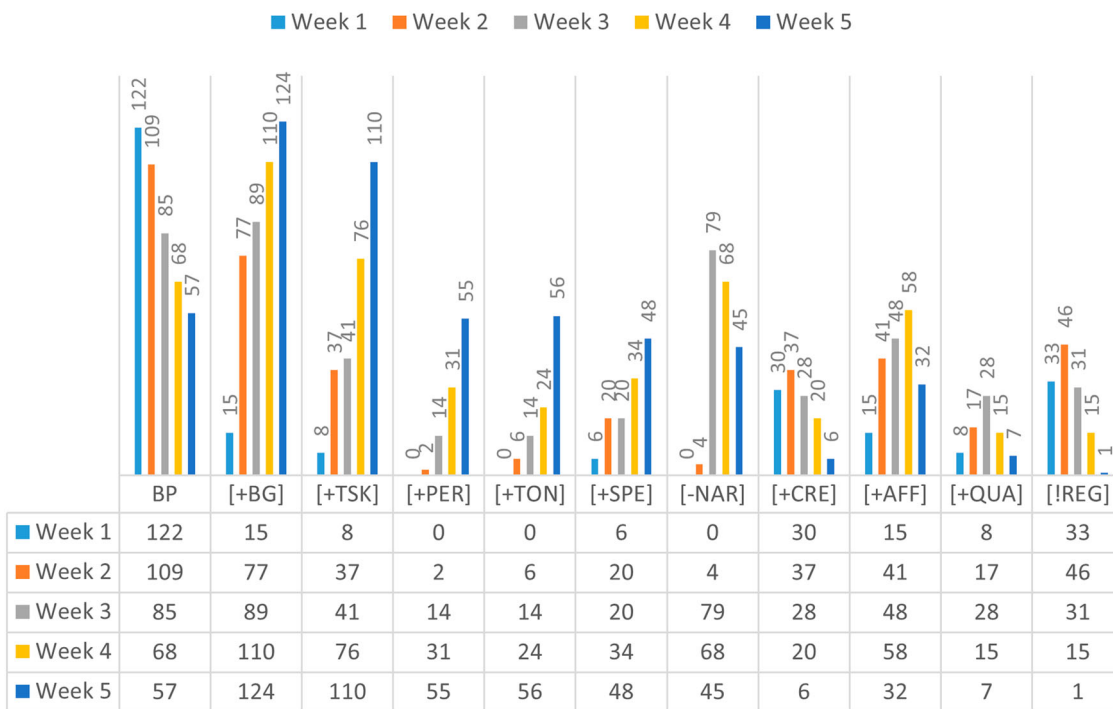
Second, the operations of revision of the four participants were gathered, coded, and categorized by the error types (see Tables 2–5 respectively for each participant). The taxonomy of errors was based on the coding instruments developed and used in the work by Ferris (2006) and Han and Hyland (2015). According to the results, ChatGPT has provided an average of 11 pieces of corrective feedback for Emma per writing task. Emma performed outstandingly with 74.55% of correct revision and actively used substitutions to correct her errors (14.55%), leaving a relatively limited amount of incorrectly executed revision (1.82%) and a low rate of rejection for correction suggestions (3.64%). Sophia’s

performance in revision execution was basically on par with Emma’s (received 12.4 pieces of corrective feedback per task), with a high correction rate (74.19%), a good percentage of substitution (19.35%), and a low rate of correction suggestion rejection (4.84%). Alternatively, Robert and Mia, who have received more than 22 pieces of corrective feedback per task, attained lower rates of correct revision (about 60%) and substitution ( $\leq 6.25\%$ ), higher rates of incorrect revision (16.5% and 12.5% respectively), correction suggestion rejection ( $\geq 6.9\%$ ), and deletion ( $>10\%$ ).

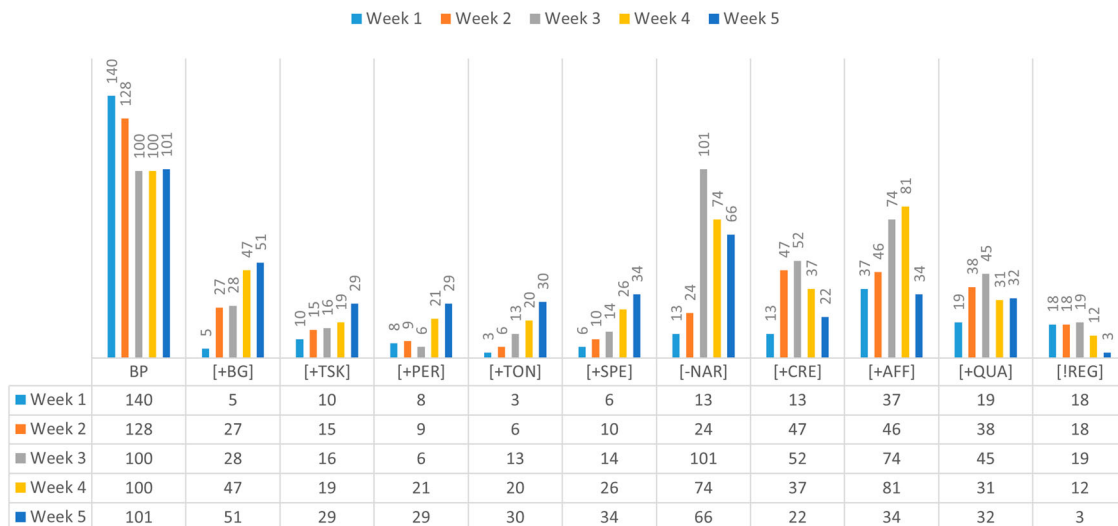
**Cognitive engagement.** The results of the LSA for the participants are displayed in Tables 6–9 respectively. In the tables, the leftmost column refers to the starting behavior, while the top row stands for the following behavior in the sequence. The behavior sequence is statistically significant when the corresponding Z value of the adjusted residual is greater than 1.96 ( $p < 0.05$ ). For example, the behavior sequence from *planning* to *feedback seeking* is statistically significant for Emma as the adjusted residual is significant ( $Z = 7.483$ ).

The above four tables were visualized diagrammatically (see Fig. 7 for the behavioral transition diagram). Each node in the diagram stands for a category of (meta)cognitive strategies, while a line linking two nodes indicates a significant behavioral transition of the sequence.

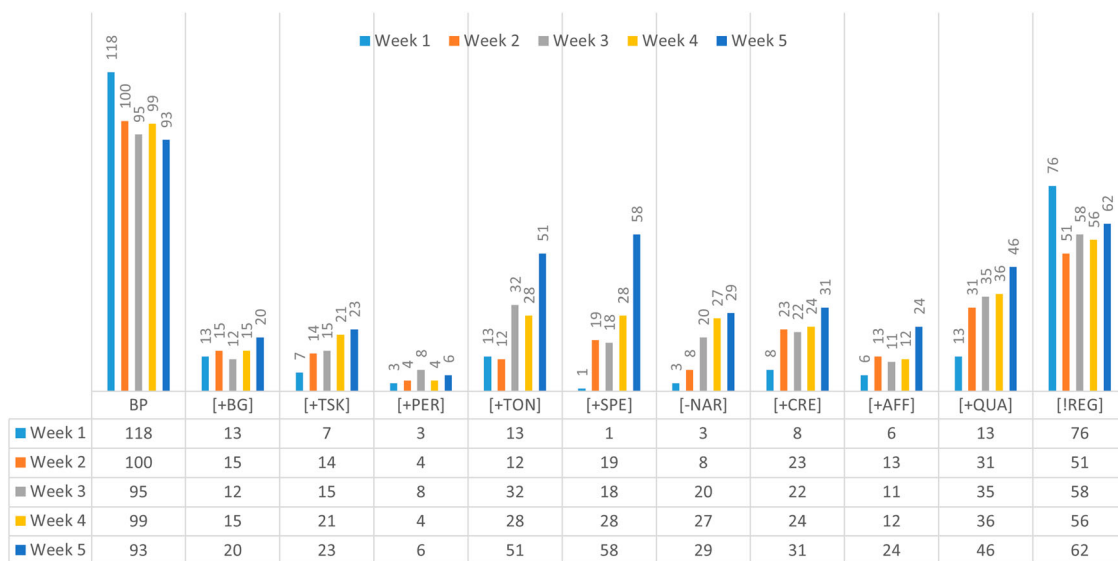
Emma has displayed a relatively higher level of metacognitive regulatory skills. The utilization of cognitive strategies to seek feedback, that is, feedback elicitation and feedback refinement, was integrated with the metacognitive regulations, i.e., monitoring and evaluation. Such integration was characterized by the bidirectional interaction between feedback seeking and metacognitive monitoring ( $Z_{F \rightarrow M} = 16.527$ ;  $Z_{M \rightarrow F} = 12.137$ ), and the similar bidirectional behavioral sequence between monitoring and feedback refinement ( $Z_{N \rightarrow M} = 9.009$ ;  $Z_{M \rightarrow N} = 12.679$ ).



**Fig. 3 Emma’s weekly actions of feedback seeking from ChatGPT.** BP: minimal prompt; [+BG]: providing background information; [+TSK]: providing task requirement; [+PER]: providing virtual persona; [+TON]: ask to feedback with ascertain style and tone; [+SPE]: with additional specific demands; [-NAR]: ask to narrow down feedback foci; [+CRE]: ask to check credibility; [+Aff]: provide affective evaluation to regenerate feedback; [+QUA]: provide quality evaluation to regenerate feedback; [IREG]: totally regenerate feedback.Credit: Based on Yan (2024b).



**Fig. 4 Sophia's weekly actions of feedback seeking from ChatGPT.** BP: minimal prompt; [+BG]: providing background information; [+TSK]: providing task requirement; [+PER]: providing virtual persona; [+TON]: ask to feedback with ascertain style and tone; [-SPE]: with additional specific demands; [-NAR]: ask to narrow down feedback foci; [+CRE]: ask to check credibility; [+Aff]: provide affective evaluation to regenerate feedback; [+QUA]: provide quality evaluation to regenerate feedback; [!REG]: totally regenerate feedback.Credit: Based on Yan (2024b).



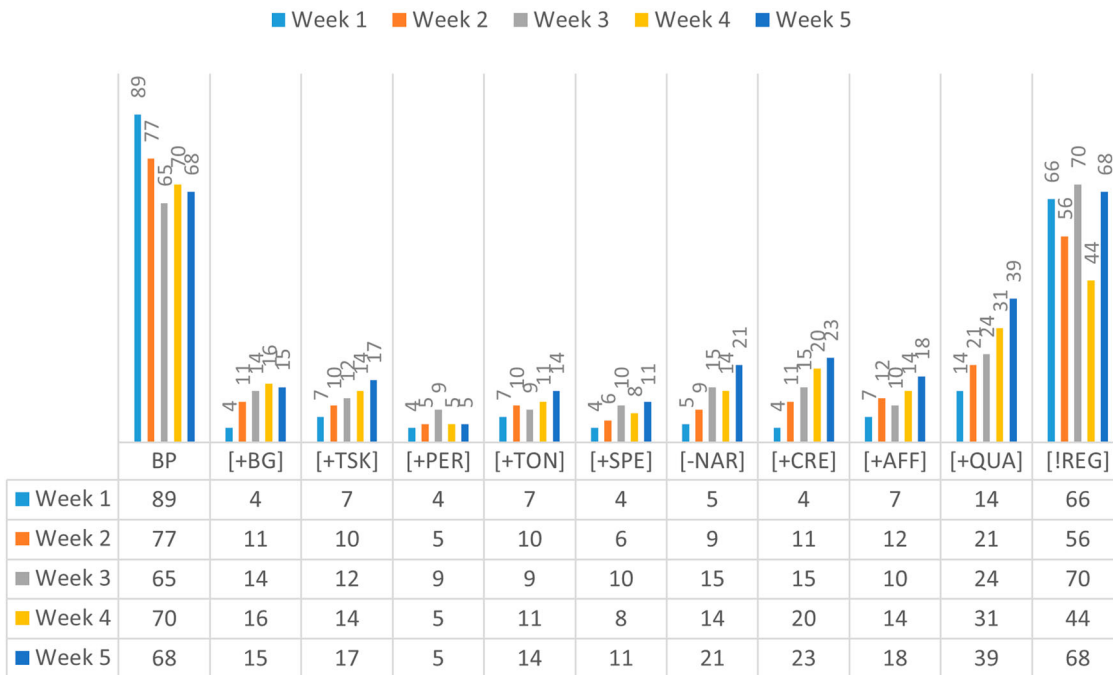
**Fig. 5 Robert's weekly actions of feedback seeking from ChatGPT.** BP: minimal prompt; [+BG]: providing background information; [+TSK]: providing task requirement; [+PER]: providing virtual persona; [+TON]: ask to feedback with ascertain style and tone; [-SPE]: with additional specific demands; [-NAR]: ask to narrow down feedback foci; [+CRE]: ask to check credibility; [+Aff]: provide affective evaluation to regenerate feedback; [+QUA]: provide quality evaluation to regenerate feedback; [!REG]: totally regenerate feedback.Credit: Based on Yan (2024b).

Sophia has demonstrated a similar pattern of utilizing cognitive and metacognitive strategies but in a relatively weaker fashion. Sharing a similar diagrammatical structure, the role of metacognitive monitoring has been reduced, typically in the feedback refinement processes (as indicated by the unidirectional sequence of  $M \rightarrow N$ ,  $Z_{M \rightarrow N} = 15.209$ ). However, the role played by metacognitive monitoring during the feedback elicitation processes remained strong (as indicated by the bidirectional behavioral sequence of  $F \rightleftharpoons M$ ,  $Z_{F \rightarrow M} = 18.15$ ;  $Z_{M \rightarrow F} = 3.834$ ).

Contrarily, the diagrams of Robert and Mia were simple and absent of the interweaving between cognitive and metacognitive strategies. In Robert's case, metacognitive strategies, i.e., monitoring and evaluation, were involved in the learning processes. He was incapable of effectively and metacognitively regulating his learning behaviors, resulting in most of his feedback elicitation

and refinement being one-off activities (as indicated by the unidirectional sequences of  $Z_{N \rightarrow M} = 15.633$ ;  $Z_{M \rightarrow E} = 15.126$ ; and  $Z_{E \rightarrow D} = 12.911$ ). Similarly, Mia has failed to integrate cognitive and metacognitive strategies. Compared to Robert, her case was even worse, as the metacognitive monitoring and evaluation were eventually severed from her feedback-seeking and revision behaviors (as indicated by  $Z_{N \rightarrow M} = 8.698$ ;  $Z_{N \rightarrow D} = 9.755$ ;  $Z_{M \rightarrow E} = 10.419$ ; and the disconnection between E and D).

**Affective engagement.** In the interview, all four participants were invited to express their affective engagement with AWCF provided by ChatGPT. We used four representative quotes to represent the four major themes that emerged from the qualitative data: (1) a beneficial journey; (2) challenges and mental



**Fig. 6 Mia's weekly actions of feedback seeking from ChatGPT.** BP: minimal prompt; [+BG]: providing background information; [+TSK]: providing task requirement; [+PER]: providing virtual persona; [+TON]: ask to feedback with ascertain style and tone; [+SPE]: with additional specific demands; [-NAR]: ask to narrow down feedback foci; [+CRE]: ask to check credibility; [+Aff]: provide affective evaluation to regenerate feedback; [+QUA]: provide quality evaluation to regenerate feedback; [!REG]: totally regenerate feedback. Credit: Based on Yan (2024b).

Error types	No.	Revision operations				
		Correct	Incorrect	Substitution	Deletion	No revision
Word choice	12	8		4		
Verb tense	3	3				
Verb form	4	3			1	
Word form						
Article	4	4				
Singular/plural						
Run-on	5	3		1	1	
Fragments	7	5	1		1	
Spelling	3	3				
Sentence structure	7	4		2		1
Phrase and idiom	5	4				1
Preposition	2	2				
Conjunctions	3	2		1		
Total (Percentage)	55 (100%)	41 (74.55%)	1 (1.82%)	8 (14.55%)	3 (5.45%)	2 (3.64%)

stresses; (3) easier to deal with GAI-generated negative feedback; and (4) continuous usage in the future.

First, students described the overall journey of using ChatGPT for AWCF as a beneficial and interesting experience. Students showcased remarkable trust in the quality of ChatGPT-generated AWCF, especially when their skills at writing prompts increased. Emma described her experiences as a “fun journey.” She was rather satisfied with ChatGPT-generated feedback, as it was of “remarkable quality and great versatility.” Sophia, sharing relatively a large proportion of Emma’s viewpoints, summarized her experiences during the project as a “thrilling journey in a bizarre yet magnificent site.” She reported that the quality of ChatGPT-generated feedback was not always stable yet mostly trustworthy and clear to follow. Robert, seeing his experiences as a “ride on the highway,” was satisfied with ChatGPT as a feedback provider

for its promptness and automated workflow. Mia concluded her journey with the project as a “shocking and slow-paced exploration.” She was satisfied with the tool and the learning environment, but not so much with her own progress.

Second, students identified the cognitive challenges they have faced and the resultant mental stresses. A consensus reached by the participants was the logistical issues, particularly the time spent seeking and refining ChatGPT’s responses while using ChatGPT for AWCF. For example, Emma reflected that the processes took her a relatively longer time and were a little bit mentally taxing, as she must “try very hard to seek better prompts that will bring feedback of higher quality and value.” Sophia expressed her desire for more training and scaffolding from teachers since one-on-one conversations with ChatGPT cannot be “sustained with fruitful outcomes.” The feedback-seeking and revision processes were “interesting, rewarding, but challenging”



**Table 3 Revision operation of Sophia.**

Error types	No.	Revision operations				
		Correct	Incorrect	Substitution	Deletion	No revision
Word choice	6	4		1		1
Verb tense	3	3				
Verb form	2	2				
Word form	1	1				
Article	3	3				
Singular/plural	2	2				
Run-on	3	2		1		
Fragments	5	3		2		
Spelling	4	4				
Sentence structure	8	6		1		1
Phrase and idiom	16	10		4	1	1
Preposition	4	3		1		
Conjunctions	5	3		2		
Total (Percentage)	62 (100%)	46 (74.19%)	0 (0.00%)	12 (19.35%)	1 (1.61%)	3 (4.84%)

**Table 4 Revision operation of Robert.**

Error types	No.	Revision operations				
		Correct	Incorrect	Substitution	Deletion	No revision
Word choice	6	4			1	1
Verb tense	7	3	1	1	1	1
Verb form	5	3	1		1	
Word form	6	3	1		2	
Article	11	5	2	1	2	1
Singular/plural	8	8				
Run-on	10	3	5	1		1
Fragments	12	4	6		2	
Spelling	8	7			1	
Sentence Structure	11	7	2		2	
Phrase and idiom	17	13	1	1		2
Preposition	8	7				1
Conjunctions	6	4		1		1
Total (Percentage)	115 (100%)	71 (61.74%)	19 (16.52%)	5 (4.35%)	12 (10.43%)	8 (6.96%)

**Table 5 Revision operation of Mia.**

Error types	No.	Revision operations				
		Correct	Incorrect	Substitution	Deletion	No revision
Word choice	7	4	1	1	1	
Verb tense	4	4				
Verb form	5	4	1			
Word form	12	8	1	2		1
Article	7	7				
Singular/plural	8	6			1	1
Run-on	6	7	2		1	2
Fragments	12	5	4	1	1	1
Spelling	8	7			1	
Sentence structure	14	10	3	2	2	1
Phrase and idiom	13	7	1	1	2	2
Preposition	9	7	1		1	
Conjunctions	7	8			2	3
Total (Percentage)	112 (100%)	68 (60.71%)	14 (12.50%)	7 (6.25%)	12 (10.71%)	11 (9.82%)

to her, and she was somewhat mentally stressed after using ChatGPT continuously for feedback. Mia explained that the feedback-seeking process was rewarding but hard and took her too much time since she regarded herself as a slow-paced learner. The only exception is Robert, who found that the feedback-

seeking processes were “a little bit boring” but not mentally taxing at all since he was confident in his digital competence.

Third, students favored ChatGPT when the tones of AWCF were negative and harsh. Compared to the traditional scenarios, the students were relieved of the shame and “losing face”

**Table 6 Lag sequential analysis of Emma (Frequencies and adjusted residuals).**

	P	M	E	F	N	D	R
P	<b>154 (14.353)*</b>	32 (-5.526)	47 (-3.965)	<b>121 (7.483)*</b>	35 (-2.165)	25 (-4.962)	41 (-4.597)
M	40 (-3.517)	34 (-5.87)	62 (-2.676)	<b>162 (12.137)*</b>	<b>131 (12.679)*</b>	34 (-4.163)	29 (-6.86)
E	36 (-4.491)	27 (-7.207)	<b>155 (8.887)*</b>	34 (-5.922)	41 (-2.143)	<b>172 (15.566)*</b>	52 (-4.253)
F	41 (-2.905)	<b>192 (16.527)*</b>	36 (-5.626)	45 (-3.505)	41 (-1.318)	56 (-0.378)	53 (-3.122)
N	51 (-1.922)	<b>144 (9.009)*</b>	<b>162 (10.512)*</b>	31 (-5.951)	27 (-4.022)	37 (-3.733)	41 (-5.277)
D	55 (0.038)	34 (-4.645)	29 (-5.831)	54 (-1.374)	35 (-1.621)	42 (-1.753)	<b>172 (14.668)*</b>
R	12 (-1.844)	7 (-3.79)	9 (-3.583)	3 (-4.565)	6 (-2.664)	11 (-1.97)	<b>101 (17.417)*</b>

\*p < 0.05, Z > 1.96

**Table 7 Lag sequential analysis of Sophia (Frequencies and adjusted residuals).**

	P	M	E	F	N	D	R
P	<b>78 (10.658)*</b>	14 (-4.847)	31 (-2.295)	<b>98 (9.788)*</b>	14 (-5.101)	12 (-3.704)	12 (-4.397)
M	12 (-3.841)	21 (-3.535)	25 (-3.464)	<b>66 (3.834)*</b>	<b>121 (15.209)*</b>	12 (-3.743)	4 (-6.096)
E	17 (-1.375)	15 (-3.242)	<b>44 (2.171)*</b>	21 (-2.754)	16 (-3.264)	<b>79 (13.9)*</b>	8 (-4.05)
F	14 (-2.285)	<b>117 (18.15)*</b>	21 (-2.798)	14 (-4.389)	22 (-2.227)	13 (-2.431)	7 (-4.444)
N	12 (-1.556)	15 (-2.049)	<b>87 (13.851)*</b>	13 (-3.179)	12 (-2.968)	9 (-2.284)	9 (-2.827)
D	10 (-1.178)	11 (-1.987)	8 (-3.194)	9 (-3.07)	16 (-0.871)	13 (-0.198)	<b>56 (11.42)*</b>
R	4 (-2.109)	2 (-3.497)	1 (-4.088)	3 (-3.595)	3 (-3.318)	6 (-1.357)	<b>70 (19.269)*</b>

\*p < 0.05, Z > 1.96

**Table 8 Lag sequential analysis of Robert (Frequencies and adjusted residuals).**

	P	M	E	F	N	D	R
P	<b>121 (11.062)*</b>	24 (-4.202)	44 (-6.649)	<b>161 (13.492)*</b>	45 (-2.821)	32 (-4.13)	12 (-6.939)
M	41 (-0.555)	32 (-1.151)	<b>177 (15.126)*</b>	21 (-5.881)	41 (-1.575)	35 (-1.884)	4 (-7.029)
E	8 (-3.932)	7 (-3.587)	<b>65 (4.099)*</b>	14 (-3.839)	9 (-4.262)	<b>84 (12.911)*</b>	15 (-2.116)
F	15 (-2.734)	21 (-0.625)	18 (-5.026)	31 (-0.912)	<b>104 (15.06)*</b>	21 (-1.599)	9 (-3.839)
N	14 (-2.619)	<b>87 (15.633)*</b>	37 (-1.122)	25 (-1.661)	18 (-2.383)	16 (-2.328)	7 (-4.005)
D	7 (-2.306)	5 (-2.433)	4 (-4.96)	7 (-3.217)	12 (-1.395)	14 (-0.442)	<b>71 (16.551)*</b>
R	3 (-3.672)	4 (-2.958)	7 (-4.582)	12 (-2.266)	9 (-2.501)	14 (-0.779)	<b>81 (18.433)*</b>

\*p < 0.05, Z > 1.96.

**Table 9 Lag sequential analysis of Mia (Frequencies and adjusted residuals).**

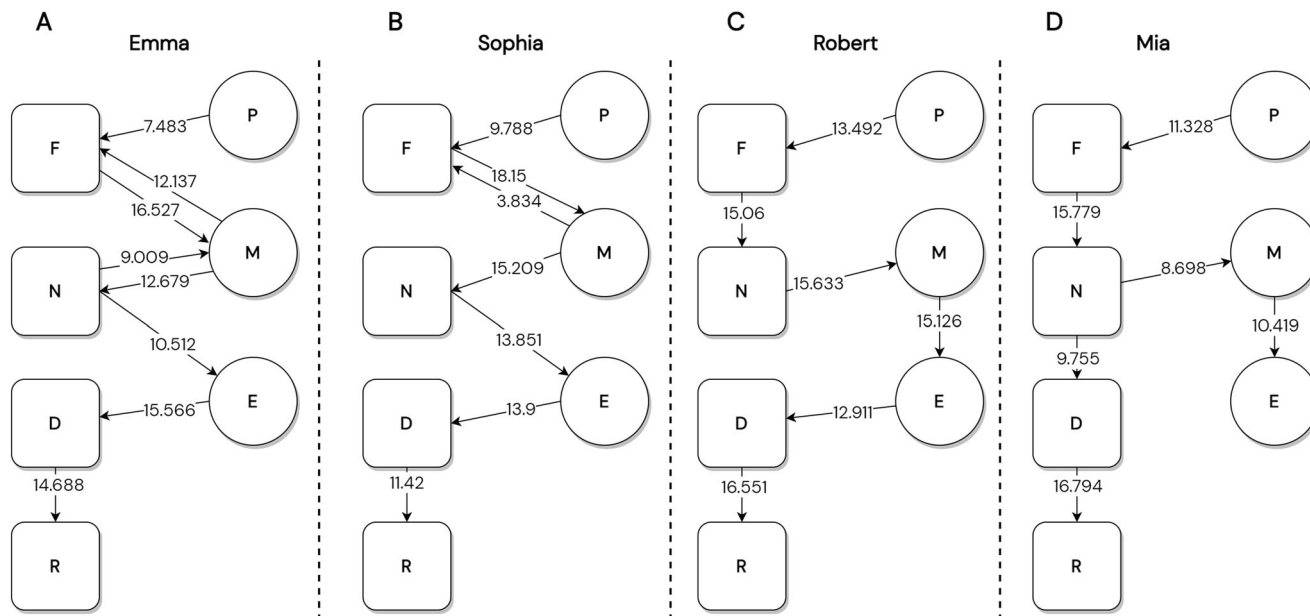
	P	M	E	F	N	D	R
P	<b>81 (13.005)*</b>	16 (-4.155)	21 (-3.282)	<b>78 (11.328)*</b>	14 (-4.242)	13 (-3.634)	6 (-5.846)
M	31 (-0.477)	<b>98 (7.898)*</b>	<b>114 (10.419)*</b>	13 (-4.586)	32 (-2.862)	17 (-4.5)	7 (-7.23)
E	14 (-2.101)	25 (-2.013)	<b>86 (9.869)*</b>	14 (-2.512)	31 (-0.463)	33 (1.006)	11 (-4.538)
F	14 (-1.865)	12 (-4.319)	16 (-3.588)	21 (-0.64)	<b>108 (15.779)*</b>	24 (-0.636)	8 (-4.904)
N	21 (-2.691)	<b>105 (8.698)*</b>	21 (-5.44)	31 (-1.286)	34 (-2.752)	<b>96 (9.755)*</b>	14 (-6.217)
D	15 (-2.773)	12 (-5.576)	16 (-4.931)	28 (-0.483)	31 (-1.732)	26 (-1.654)	<b>131 (16.794)*</b>
R	4 (-3.233)	9 (-3.48)	7 (-4.006)	12 (-1.341)	12 (-2.531)	14 (-1.333)	<b>87 (15.355)*</b>

\*p < 0.05, Z > 1.96

experience in front of teachers and peers. Emma asserted that it's much easier for her to accept negative feedback from AI systems than teachers in the classroom. Mia shared a similar feeling that handling ChatGPT-generated negative feedback feels like those from an anonymous agent.

Finally, students expressed their interest in continuously using ChatGPT in the future. On a broader spectrum, the students acknowledge the value and applicability of ChatGPT as an AWCf provider. As Emma remarked, "using AI for

corrective feedback will be normal in the future, and the tips and tricks we have explored will be of valuable significance". Sophia was sure that she would continue to explore the more advanced features of ChatGPT in writing classrooms, but Mia was worried that she would be outperformed by her classmates as she was slow to pick up the more sophisticated tricks and usage. Robert claimed rather straightforwardly that he would be using ChatGPT after the project to "avoid face-to-face feedback from teachers".



**Fig. 7 Behavioral transition diagram of the participants.** P: planning, referring to allocation of time, resources for the following-up feedback and writing processes; M: monitoring, referring to an on-going process in which the quality feedback is observed and compared; E: evaluation, referring to an appraisal of the value and cost for a potential revision or correction based on the feedback selected from the monitoring process; F: feedback elicitation, referring to using the interactive communication with ChatGPT to elicit AWCF; N: feedback refinement, referring to comparing and finalizing potential feedback and ask ChatGPT to regenerate for quality improvement if the quality is unsatisfactory; D: making decision, referring to a final appraisal of the feedback quality and translate the feedback to a potential revision; R: executing the revision, referring to applying the finalized revision to the writing products.

**Discussions**

The study explored the behavioral, cognitive, and affective engagement of L2 writers with corrective feedback provided by ChatGPT in feedback-seeking and revision processes. The findings are categorically presented and discussed in the following sections against existing research and theoretical insights.

**Proactive feedback seeking and revision behaviors.** The four participants’ behavioral engagement revealed that students were actively involved in the feedback-seeking and revision execution processes. At first glance, all four participants have made progress in seeking feedback from ChatGPT throughout the weeks. Internally, high language-proficiency learners (represented by Emma and Sophia), showed a more sophisticated approach to refining ChatGPT prompts. Instead of repeatedly asking ChatGPT to regenerate feedback, the two learners focused on the quality and content richness of the prompts. The observed varieties could be explained by the process of *inner feedback*, a term advocated by Nicol (2021) to represent the natural processing and comparison after learners’ exposure to feedback. Based on the findings, we could infer that the ability to internally process the received feedback during the feedback seeking from ChatGPT depended on the students’ language proficiencies. From another perspective, students’ feedback-seeking behaviors revealed that students with a higher level of technological competence were likely to make more attempts in feedback elicitation and refinement. The specific result was in line with the widely accepted viewpoint that a higher level of ICT competence or digital literacy would lead to more advanced learning outcomes in a technology-enhanced learning environment (Park and Weng 2020; Yan and Wang 2022).

Similarly, participants with different language proficiencies manifested varied patterns in translating the received feedback to revision execution. Apart from the differences in total errors detected by the AI system per writing task, the most drastic discrepancies among the four participants in the revision

operation were the rate of correct revision and adoption of revision strategies. On the one hand, the rate of correct revision was higher than that from precedent research with Grammarly as a feedback provider (i.e., Koltovskaia 2020). This could be explained by the alleged strength of ChatGPT in correcting grammatical errors (H. Wu et al. 2023). On the other hand, the observation that high-proficiency language learners would make significantly more substitutions than low-proficiency learners echoed the findings of Barkaoui (2016). However, in contrast with Barkaoui’s (2016) study, low-proficiency language learners made significantly more revision deletions than their peers. Comprehensively, the students, especially the low-proficiency ones, have ineffectively utilized the corrective feedback provided by ChatGPT. This phenomenon was in line with previous literature (Warschauer and Grimes 2008; Chapelle et al. 2015).

**Diversified metacognitive regulatory skills.** Cognitively, the extent to which the participants were engaged with the ChatGPT-generated corrective feedback diversified distinctly. Generally, the students performed unsatisfactorily to metacognitively regulate their learning, especially during the feedback-seeking processes. This phenomenon was in unison with Koltovskaia’s (2020) study, where the participants failed to process AWCF effectively. Furthermore, the relatively poor metacognitive strategy use also testified to the finding of Zhang and Zhang (2022) that the AWCF hindered students’ active utilization of monitoring and evaluation strategies. Specifically, higher proficiency learners (represented by Emma and Sophia) have effectively utilized metacognitive monitoring and evaluation of the quality of the received feedback to make full use of the strength of ChatGPT; conversely, the lower proficiency learners (i.e., Robert and Mia) could not effectively integrate the metacognitive strategies with the cognitive processes. The variations in the metacognitive regulatory skills among the participants could be attributed to the view of Zheng and Yu (2018) that insufficient language

proficiency would hinder learners' ability to process feedback and revision.

Unexpected findings emerged from the comparison of the LSA results between Robert and Mia. Based on the data and the visualization, we could posit that students possessing better technological competence could compensate for their limited abilities to monitor and evaluate the quality of received feedback with intensive communication with AI systems. Such inference underlined the revolutionary affordance of ChatGPT's conversational AI system in providing a highly customizable and learner-aware environment that satisfies learners' needs through repeated and creative prompt writing (Ranalli 2018; Oppenlaender et al. 2023; Rudolph et al. 2023). Additionally, the finding was in tandem with the meta-analysis results of Wu and Yu (2023) that AI chatbots were impactful on learning outcomes. The insights would create a new understanding of students' feedback processing in a learning environment equipped with GAI-based or conversational tools.

**An affectively engaging learning environment.** The attitudinal and emotional responses towards ChatGPT-generated AWCF and the new GAI-powered learning environment were mostly positive. The overall satisfaction with and acceptance of ChatGPT as a corrective feedback provider was in line with relevant studies in the field of AWCF (Dikli and Bleyle 2014; Koltovskaia 2020). Furthermore, participants have agreed that the quality of ChatGPT-generated corrective feedback was reliable and accurate. Compared to previous research on the acceptance and evaluation of AWE systems and tools such as Grammarly, the performance of ChatGPT was convincing and well acclaimed by its users (Zhang 2017; Koltovskaia 2020; Ranalli 2021). This phenomenon could be attributed to the interplay of the computational might of the AI system (Fang et al. 2023; Wu et al. 2023) and its interactive human-machine interface (Oppenlaender et al. 2023).

However, participants stressed the mental effort expenditure that resulted from using ChatGPT in L2 writing classrooms. This was not unexpected, as AWCF providers or AWE systems have always been linked with cognitive overload in previous studies (Ranalli 2018; Barrot 2021). Nevertheless, the cognitive burden experienced by users of ChatGPT was the aggregate of mental effort expenditure for both feedback seeking and feedback processing. The finding ushered in new insights that would expand our understanding of students' cognitive load in utilizing feedback for L2 writing. Moreover, the finding was in tandem with a recent research trend beyond the scope of AWCF studies to explore how to effectively compose high-quality ChatGPT prompts (Oppenlaender et al. 2023; White et al. 2023) and how to develop students' abilities to communicate with GAI systems (Yan 2023; Yan 2024b).

## Conclusion

The multiple mixed-method case study, involving four students with different language proficiencies and technological competences from an EFL program, has explored L2 writers' engagement with ChatGPT-provided corrective feedback from behavioral, cognitive, and affective perspectives. The findings revealed that: 1) students were behaviorally engaged with ChatGPT-generated feedback; however, their feedback-seeking behaviors and revision operations are highly related to language proficiencies and technological competences; 2) only high language proficiency learners could cognitively engage with ChatGPT-generated AWCF by effectively utilizing metacognitive regulatory strategies; and 3) ChatGPT was well-received by participants as a powerful and affectively engaging AWCF provider.

Adding to the body of literature on students' engagement with AWCF, the study also focuses on the changes in learning brought

about by the appearance of ChatGPT. Noticeably, the research underlines the importance of technological competence for L2 learners exposed to technology-enhanced learning environments. Furthermore, as an initial effort to investigate the patterns of learning behaviors and utilization of (meta)cognitive strategies of L2 writers in a GAI-powered environment, the study offers insights into how students are involved in seeking feedback instead of receiving feedback from AWCF providers and how the feedback processing and revision processes are regulated metacognitively.

The diversity of student engagement with ChatGPT-generated corrective feedback, as manifested by the study, has significant pedagogical implications. First, ChatGPT was not only a powerful rival to its precedents but also an affectively engaging solution with which a new learning environment could be constructed. As a result, the inclusion of GAI-based applications as learning assistants in L2 classrooms should be popularized. Second, teacher scaffolding or instruction on the utilization of ChatGPT for the purposes of L2 writing pedagogy or assessment should be developed and provided. As reflected in the study, learners' individual ability to metacognitively regulate feedback seeking and revision execution is a far cry from perfection. Hence, support from instructors and peer learners is highly expected. Third, a more rational attitude towards the position of GAI-based products in education should be upheld. Instead of a "silver bullet" or a terminator of education, ChatGPT's integration in classrooms needs the enhancement of students' and instructors' multi-competence and the corresponding restructuring of instructional patterns. Finally, from an L2 learner perspective, the relatively high drop-out rate during the participant recruitment showed that, at least at the current stage, students didn't possess sufficient AI competence and domain knowledge to effectively utilize GAI for longer-time learning improvement. Thus, sustained efforts should be provided in training students of the contemporary era into better users of state-of-the-art technologies.

The study was not without limitations. First, the study adopted a multiple-case study approach methodologically. Hence, researchers should be cautious when translating or generalizing the findings of the present study to different research settings with larger populations. In follow-up research, alternative research methods could be considered to comprehensively investigate the impact of ChatGPT on a larger number of language learners. Second, the duration of the research is limited. In a five-week project, students have completed merely five writing tasks with limited exposure to ChatGPT. In subsequent studies, researchers could try to conduct longitudinal investigations through which the long-term effects of ChatGPT on the learning behaviors and outcomes of L2 learners could be uncovered. Third, the modes of sources of feedback are limited. The study partially adopted a self-regulated learning style for the participants. Hence, the role of peer learners and instructors in processing the feedback was not examined. In successive inquiries, researchers could introduce collaborative learning or peer scaffolding into the learning environment. Fourth, the impact of ChatGPT-generated feedback on writing of different genres was not studied. In future studies, researchers could delve into the effects of the AWCF provided by ChatGPT on multiple types and genres of writing. In general, with the exhibited potential of ChatGPT as a game changer for language education, the researcher hopes the study could kindle more in-depth insights into the pedagogical practice of utilizing GAI-based applications in L2 classrooms.

## Data availability

The pseudonymized data that support the findings of this study are available on request from the corresponding author. The raw

data are not publicly available due to the concern that they might disclose the privacy of the participants.

Received: 27 February 2024; Accepted: 31 July 2024;

Published online: 26 August 2024

## References

- Adams G (2019) A narrative study of the experience of feedback on a professional doctorate: 'a kind of flowing conversation'. *Stud Contin Educ* 41(2):191–206. <https://doi.org/10.1080/0158037X.2018.1526782>
- Bai L, Hu G (2017) In the face of fallible AWE feedback: how do students respond? *Educ Psychol* 37(1):67–81. <https://doi.org/10.1080/01443410.2016.1223275>
- Bakeman R, Quera V (2011) *Sequential analysis and observational methods for the behavioral sciences*. Cambridge University Press, Cambridge. <https://doi.org/10.1017/CBO9781139017343>
- Barkaoui K (2016) What and when second-language learners revise when responding to timed writing tasks on the computer: the roles of task type, second language proficiency, and keyboarding skills. *Mod Lang J* 100(1):320–340. <https://doi.org/10.1111/modl.12316>
- Barrot JS (2021) Using automated written corrective feedback in the writing classrooms: effects on L2 writing accuracy. *Comput Assist Lang Learn*. <https://doi.org/10.1080/09588221.2021.1936071>
- Bowen GA (2009) Document analysis as a qualitative research method. *Qual Res J* 9(2):27–40. <https://doi.org/10.3316/QRJ0902027>
- Braun V, Clarke V (2012) Thematic analysis. In: *APA handbook of research methods in psychology, vol 2: research designs: quantitative, qualitative, neuropsychological, and biological*. APA handbooks in psychology®. American Psychological Association, Washington, DC, pp 57–71. <https://doi.org/10.1037/13620-004>
- Carter N, Bryant-Lukosius D, DiCenso A et al. (2014) The use of triangulation in qualitative research. *Oncol Nurs Forum* 41(5):545–547. <https://doi.org/10.1188/14.ONF.545-547>
- Chapelle CA, Cotos E, Lee J (2015) Validity arguments for diagnostic assessment using automated writing evaluation. *Lang Test* 32(3):385–405. <https://doi.org/10.1177/0265532214565386>
- Creswell JW, Plano Clark VL (2018) *Designing and conducting mixed methods research*, 3rd edn. SAGE, LA
- Dikli S, Bleye S (2014) Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assess Writ* 22:1–17. <https://doi.org/10.1016/j.asw.2014.03.006>
- Dizon G, Gayed J (2021) Examining the impact of Grammarly on the quality of mobile L2 writing. *JALT CALL J* 17(2):74–92. <https://doi.org/10.29140/jaltcall.v17n2.336>
- Doyle S (2007) Member checking with older women: a framework for negotiating meaning. *Health Care Women Int* 28(10):888–908. <https://doi.org/10.1080/07399330701615325>
- Duff P (2010) *Case study research in applied linguistics*. Second language acquisition research. Routledge, New York
- Ellis R (2010) A framework for investigating oral and written corrective feedback. *Stud Second Lang Acq* 32(2):335–349. <https://doi.org/10.1017/S0272263109990544>
- Fan Y, Xu J (2020) Exploring student engagement with peer feedback on L2 writing. *J Second Lang Writ* 50:100775. <https://doi.org/10.1016/j.jslw.2020.100775>
- Fang T, Yang S, Lan K et al. (2023) Is ChatGPT a highly fluent grammatical error correction system? A comprehensive evaluation. *arXiv*. <https://doi.org/10.48550/ARXIV.2304.01746>
- Ferris D (2006) Does error feedback help student writers? New evidence on the short- and long-term effects of written error correction. In: Hyland, F, Hyland, K (eds) *Feedback in second language writing: contexts and issues*. Cambridge applied linguistics. Cambridge University Press, Cambridge, pp 81–104. <https://doi.org/10.1017/CBO9781139524742.007>
- Fleckenstein J, Leucht M, Köller O (2018) Teachers' judgement accuracy concerning CEFR levels of prospective university students. *Lang Assess Q* 15(1):90–101. <https://doi.org/10.1080/15434303.2017.1421956>
- Fleckenstein J, Liebenow LW, Meyer J (2023) Automated feedback and writing: a multi-level meta-analysis of effects on students' performance. *Front Artif Intell*. <https://doi.org/10.3389/frai.2023.1162454>
- Fu Q-K, Zou D, Xie H et al. (2022) A review of AWE feedback: types, learning outcomes, and implications. *Comput Assist Lang Learn*. <https://doi.org/10.1080/09588221.2022.2033787>
- Gong H, Yan D (2023) The impact of danmaku-based and synchronous peer feedback on L2 oral performance: a mixed-method investigation. *PLoS ONE* 18(4):e0284843. <https://doi.org/10.1371/journal.pone.0284843>
- Han Y, Gao X (2021) Research on learner engagement with written (corrective) feedback: insights and issues. In: Mercer, S, Hiver, P, Al-Hoorie, AH (eds) *Student engagement in the language classroom*. Multilingual Matters, pp 56–74. <https://doi.org/10.21832/9781788923613-007>
- Han Y, Hyland F (2015) Exploring learner engagement with written corrective feedback in a Chinese tertiary EFL classroom. *J Second Lang Writ* 30:31–44. <https://doi.org/10.1016/j.jslw.2015.08.002>
- Hiver P, Al-Hoorie AH, Vitta JP et al. (2021) Engagement in language learning: a systematic review of 20 years of research methods and definitions. *Lang Teach Res*. <https://doi.org/10.1177/13621688211001289>
- Hyland K, Hyland F (2019) Contexts and issues in feedback on L2 writing. In: Hyland, F (ed) *Feedback in second language writing: contexts and issues*, 2nd edn. Cambridge applied linguistics. Cambridge University Press, Cambridge, pp 1–22. <https://doi.org/10.1017/9781108635547.003>
- Jamshed S (2014) Qualitative research method-interviewing and observation. *J Basic Clin Pharm* 5(4):87–88. <https://doi.org/10.4103/0976-0105.141942>
- Jansen T, Vögelin C, Machts N et al. (2021) Judgment accuracy in experienced versus student teachers: assessing essays in English as a foreign language. *Teach Teach Educ* 97:103216. <https://doi.org/10.1016/j.tate.2020.103216>
- Jiao W, Wang W, Huang J et al. (2023) Is ChatGPT a good translator? Yes with GPT-4 As the engine. *arXiv*. <https://doi.org/10.48550/arXiv.2301.08745>
- Jin Y, Fan J (2011) Test for English majors (TEM) in China. *Lang Test* 28(4):589–596. <https://doi.org/10.1177/0265532211414852>
- Koltovskaia S (2020) Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: a multiple case study. *Assess Writ* 44:100450. <https://doi.org/10.1016/j.asw.2020.100450>
- Koltovskaia S, Mahapatra S (2022) Student engagement with computer-mediated teacher written corrective feedback: a case study. *JALT CALL J* 18(2):286–315. <https://doi.org/10.29140/jaltcall.v18n2.519>
- Lee H (2023) The rise of ChatGPT: exploring its potential in medical education. *Anat Sci Educ*. <https://doi.org/10.1002/ase.2270>
- Li J, Link S, Hegelheimer V (2015) Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *J Second Lang Writ* 27:1–18. <https://doi.org/10.1016/j.jslw.2014.10.004>
- Liu S, Yu G (2022) L2 learners' engagement with automated feedback: an eye-tracking study. *Lang Learn Technol* 26(2):78–105. 10125/73480
- Mercer S (2019) Language learner engagement: setting the scene. In: Gao, X (ed) *Second handbook of English language teaching*. Springer international handbooks of education, Springer International Publishing, Cham, pp 1–19. [https://doi.org/10.1007/978-3-319-58542-0\\_40-1](https://doi.org/10.1007/978-3-319-58542-0_40-1)
- Mizumoto A, Eguchi M (2023) Exploring the potential of using an AI language model for automated essay scoring. *Res Methods Appl Linguist* 2(2):100050. <https://doi.org/10.1016/j.rmal.2023.100050>
- Naamati-Schneider L, Alt D (2024) Beyond digital literacy: the era of AI-powered assistants and evolving user skills. *Educ Inf Technol*. <https://doi.org/10.1007/s10639-024-12694-z>
- Nelson MM, Schunn CD (2009) The nature of feedback: how different types of peer feedback affect writing performance. *Instr Sci* 37(4):375–401. <https://doi.org/10.1007/s11251-008-9053-x>
- Nicol D (2011) The power of internal feedback: exploiting natural comparison processes. *Assess Eval High Educ* 46(5):756–778. <https://doi.org/10.1080/02602938.2020.1823314>
- O'Neill R, Russell A (2019) Stop! Grammar time: university students' perceptions of the automated feedback program Grammarly. *Australas J Educ Technol*. <https://doi.org/10.14742/ajet.3795>
- Oppenlaender J, Linder R, Silvennoinen J (2023) Prompting AI art: an investigation into the creative skill of prompt engineering. <https://doi.org/10.48550/arXiv.2303.13534>
- Palinkas LA, Horwitz SM, Green CA et al. (2015) Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Adm Policy Ment Health* 42(5):533–544. <https://doi.org/10.1007/s10488-013-0528-y>
- Park S, Weng W (2020) The relationship between ICT-related factors and student academic achievement and the moderating effect of country economic index across 39 countries: using multilevel structural equation modelling. *Educ Technol Soc* 23(3):1–15
- Pohl M, Wallner G, Kriglstein S (2016) Using lag-sequential analysis for understanding interaction sequences in visualizations. *Int J Hum Comput Stud* 96:54–66. <https://doi.org/10.1016/j.ijhcs.2016.07.006>
- Rad HS, Alipour R, Jafarpour A (2023) Using artificial intelligence to foster students' writing feedback literacy, engagement, and outcome: a case of Wordtune application. *Interact Learn Environ*. <https://doi.org/10.1080/10494820.2023.2208170>
- Ranalli J (2018) Automated written corrective feedback: how well can students make use of it? *Comput Assist Lang Learn* 31(7):653–674. <https://doi.org/10.1080/09588221.2018.1428994>
- Ranalli J (2021) L2 student engagement with automated feedback on writing: potential for learning and issues of trust. *J Second Lang Writ* 52:100816. <https://doi.org/10.1016/j.jslw.2021.100816>

- Rudolph J, Tan S, Tan S (2023) ChatGPT: bullshit spewer or the end of traditional assessments in higher education? *J Appl Learn Teach* 6(1):1–22. <https://doi.org/10.37074/jalt.2023.6.1.9>
- Shi Y (2021) Exploring learner engagement with multiple sources of feedback on L2 writing across genres. *Front. Psychol.* <https://doi.org/10.3389/fpsyg.2021.758867>
- Sonnenberg C, Bannert M (2015) Discovering the effects of metacognitive prompts on the sequential structure of SRL-processes using process mining techniques. *J Learn Anal* 2(1):72–100. <https://doi.org/10.18608/jla.2015.21.5>
- Stake RE (1995) *The art of case study research*. Sage Publications, Thousand Oaks
- Steiss J, Tate T, Graham S et al. (2024) Comparing the quality of human and ChatGPT feedback of students' writing. *Eur Res Int* 91:101894. <https://doi.org/10.1016/j.learninstruc.2024.101894>
- Stevenson M, Phakiti A (2019) Automated feedback and second language writing. In: Hyland, F, Hyland, K (eds) *Feedback in second language writing: contexts and issues*, 2nd edn. Cambridge applied linguistics. Cambridge University Press, Cambridge, pp 125–142. <https://doi.org/10.1017/9781108635547.009>
- Tan S, Cho YW, Xu W (2022) Exploring the effects of automated written corrective feedback, computer-mediated peer feedback and their combination mode on EFL learner's writing performance. *Interact Learn Environ.* <https://doi.org/10.1080/10494820.2022.2066137>
- Tonmoy SMTI, Zaman SMM, Jain V et al. (2024) A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv.* <https://doi.org/10.48550/arXiv.2401.01313>
- Tseng W, Warschauer M (2023) AI-writing tools in education: if you can't beat them, join them. *J China Comput Assist Lang Learn.* <https://doi.org/10.1515/jccall-2023-0008>
- Warschauer M, Grimes D (2008) Automated writing assessment in the classroom. *Pedagogies* 3(1):22–36. <https://doi.org/10.1080/15544800701771580>
- Watkins MW, Pacheco M (2000) Interobserver agreement in behavioral research: Importance and calculation. *J Behav Educ* 10(4):205–212. <https://doi.org/10.1023/A:1012295615144>
- White J, Fu Q, Hays S et al. (2023) A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv.* <https://doi.org/10.48550/ARXIV.2302.11382>
- Wood J (2022) Supporting the uptake process with dialogic peer screencast feedback: a sociomaterial perspective. *Teach Higher Educ.* <https://doi.org/10.1080/13562517.2022.2042243>
- Wu H, Wang W, Wan Y et al. (2023) ChatGPT or Grammarly? Evaluating ChatGPT on grammatical error correction benchmark. *arXiv.* <https://doi.org/10.48550/ARXIV.2303.13648>
- Wu R, Yu Z (2023) Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis. *Brit J Educ Technol.* <https://doi.org/10.1111/bjet.13334>
- Yan D (2023) Impact of ChatGPT on learners in a L2 writing practicum: an exploratory investigation. *Educ Inf Technol* 28(11):13943–13967. <https://doi.org/10.1007/s10639-023-11742-4>
- Yan D (2024a) Rubric co-creation to promote quality, interactivity and uptake of peer feedback. *Assess Eval Higher Educ.* <https://doi.org/10.1080/02602938.2024.2333005>
- Yan D (2024b) Feedback seeking abilities of L2 writers using ChatGPT: a mixed method multiple case study. *Kybernetes.* <https://doi.org/10.1108/K-09-2023-1933>
- Yan D, Wang J (2022) Teaching data science to undergraduate translation trainees: pilot evaluation of a task-based course. *Front Psychol* 13:939689. <https://doi.org/10.3389/fpsyg.2022.939689>
- Yin RK (2013) *Case study research: design and methods*. 5th edn. SAGE Publications, Los Angeles
- Zhang J, Zhang LJ (2022) The effect of feedback on metacognitive strategy use in EFL writing. *Comput Assist Lang Learn.* <https://doi.org/10.1080/09588221.2022.2069822>
- Zhang Z (2017) Student engagement with computer-generated feedback: a case study. *ELT J* 71(3):317–328. <https://doi.org/10.1093/elt/ccw089>
- Zhang Z, Hyland K (2018) Student engagement with teacher and automated feedback on L2 writing. *Assess Writ* 36:90–102. <https://doi.org/10.1016/j.asw.2018.02.004>
- Zhang Z, Hyland K (2023) Student engagement with peer feedback in L2 writing: Insights from reflective journaling and revising practices. *Assess Writ* 58:100784. <https://doi.org/10.1016/j.asw.2023.100784>
- Zheng L, Niu J, Zhong L et al. (2021) The effectiveness of artificial intelligence on learning achievement and learning perception: a meta-analysis. *Interact Learn Environ.* <https://doi.org/10.1080/10494820.2021.2015693>
- Zheng Y, Yu S (2018) Student engagement with teacher written corrective feedback in EFL writing: a case study of Chinese lower-proficiency students. *Assess Writ* 37:13–24. <https://doi.org/10.1016/j.asw.2018.03.001>

## Acknowledgements

This research project is supported by funding from the Young Researcher Program of Xinyang Agriculture and Forestry University [Grant QN2022049, QN2021033]. We would also like to thank all the anonymous reviewers for the constructive feedback.

## Author contributions

Da Yan: conceptualization, data curation, writing—original draft, formal analysis, project administration, writing—review, and editing. Shuxian Zhang: data curation, coding, and writing—review.

## Competing interests

The authors declare no competing interests.

## Ethical approval

At the time of the study, Xinyang Agriculture and Forestry University had no policy for ethical clearance, nor did it have an ethical committee. Thus, ethical approval was obtained from the School of Foreign Languages, Xinyang Agriculture and Forestry University in December 2022. The study was conducted in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards, including existing laws and regulations on personal data, privacy, and research data management.

## Informed consent

Written informed consent was obtained via email from all participants before the study in December 2022.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1057/s41599-024-03543-y>.

**Correspondence** and requests for materials should be addressed to Shuxian Zhang.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024