

Shaofeng Li\*

# Generative AI and Second Language Writing

<https://doi.org/10.1515/dsl-2025-0007>

Received April 22, 2025; accepted May 11, 2025; published online June 11, 2025

**Abstract:** This article provides a critical synthesis and analysis of the research on the application of generative AI (GenAI) in second language (L2) writing. It conceptualizes GenAI literacy, synthesizes the research on written feedback, establishes a framework for prompt engineering, critiques the research examining the validity of GenAI ratings in writing assessment, and summarizes empirical evidence on the differences between GenAI and human writing. Specifically, the following findings and arguments are presented and discussed. GenAI literacy consists of four components pertaining to users' competence and knowledge of GenAI basics, effective use, output evaluation, and ethics. The research on written feedback shows that teacher feedback focuses more on content, while GenAI feedback focuses more on organization. This research also suggests a need for criteria-based feedback and feedback evaluation. Prompt engineering is discussed along three dimensions: input, task, and output, followed by snapshots of prompts used in feedback research. The studies on writing assessment reveal that GenAI ratings are more consistent with human ratings when GenAI is trained using a large number of scored essays and when the rating criteria are well-defined. Comparisons of GenAI and human writing demonstrate that GenAI writing is more formal, academic, and impersonal, while human writing is more personal, creative, and linguistically accessible. This article concludes by making sense of the research findings, identifying future directions, and proposing three principles that may guide the research, practice, and theory construction for GenAI: individualization, domain-specificity, and writer agency.

**Keywords:** generative AI; ChatGPT; second language writing; corrective feedback; AI literacy; writing assessment

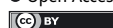
## 1 Introduction

With the launch of ChatGPT in November 2022, generative AI (GenAI) has ushered in a new era in human history. GenAI, represented by ChatGPT, DeepSeek, and other programs with similar functions, is a technology that leverages patterns extracted

---

\*Corresponding author: Shaofeng Li, The Hong Kong Polytechnic University, Kowloon, Hong Kong, E-mail: shaofeng.li@polyu.edu.hk

 Open Access. © 2025 the author(s), published by De Gruyter on behalf of Chongqing University, China

 This work is licensed under the Creative Commons Attribution 4.0 International License.

from training materials to generate responses to human inquiries by predicting the most likely next words. GenAI has the following characteristics. First, it is a large *language* model based on linguistic probabilities; therefore, its outputs are determined by linguistic contingencies and the strengths of bonds between linguistic units. Because it is a language model and is not based on contextual, rhetorical, pragmatic, or semantic cues, it may generate responses that are linguistically coherent but semantically illogical and unreasonable. Second, it is generative in that it can produce new information based on existing input materials. The input materials are publicly available materials selected and fed by IT specialists, engineers, or personnel whose qualifications are unknown. The selection criteria for training materials are also unknown. The latest update to the training materials for ChatGPT, probably the most popular GenAI tool, occurred in October 2023. It is important to clarify that GenAI is not connected with any external source of data, nor the Internet, so there is no ongoing learning. It is also necessary to clarify that according to ChatGPT, user responses are not used for training purposes, which may help dispel qualms and concerns over privacy and confidentiality. Third, it is interactive and can engage in human-like conversations with users. It may generate multiple turns, keep previous turns in memory, and retrieve and process previously stored information when requested. However, the kind of interaction it engages in with users lacks key features of natural conversations (Voss and Waring 2024). For example, it is not sensitive to user characteristics and does not use appropriate communication strategies in response to user differences; it also does not make constant adjustments based on the flow of the conversation and ongoing dynamics. Fourth, it simulates human intelligence and has the ability to perform various higher-order tasks that require or involve information processing such as analyzing, categorizing, comparing, inferencing, reasoning, summarizing, and so on.

GenAI's versatility is evident in second language (L2) writing. Writers can use GenAI to brainstorm ideas, create outlines, summarize source information, translate information in their first language into the second language, search for L2 forms that match planned messages, proofread their essays for language, content, and rhetoric, etc. Teachers can utilize GenAI to evaluate and provide feedback on students' writing, which is logistically challenging given the large number of assignments teachers have to evaluate. Other issues such as teachers' failure to recognize errors and provide accurate, consistent feedback, as well as struggles teachers experience in providing feedback on the organizational aspects of writing (Lee 2008; Li and Vuono 2019; Truscott 1996), can be easily addressed with the assistance of GenAI. In addition to benefiting students and teachers, GenAI has brought opportunities to researchers, whose research territory and terrain have been expanded and reshaped. Despite

GenAI's affordances and opportunities for L2 writing, there is a lack of guidance on its effective use, there are pitfalls that have been identified by research, and many pressing issues remain unexplored in the research. This article synthesizes the various strands of research investigating the role of GenAI in second language writing, including GenAI literacy, feedback on L2 writing, prompt engineering, discourse comparison between GenAI and human writing, and ethics. Integrated into the synthesis of research findings is a discussion of new concepts and the implications of the research for theory construction, further research, and L2 writing practice and pedagogy.

## 2 GenAI Literacy

GenAI literacy refers to users' knowledge of the fundamentals of GenAI's mechanism, ethics, affordances, and limitations, and users' ability to effectively use GenAI. GenAI literacy is an important concept due to GenAI's ubiquity and limitations and users' lack of knowledge about how to maximize the benefits of this new technological tool. It is important to clarify that GenAI literacy involves not only users' awareness of GenAI's limitations (bias, inaccurate output, etc.), but also users' understanding of its functions and strategies to leverage it effectively. GenAI literacy entails both declarative knowledge – knowledge about its operational principles, affordances, etc. – and also procedural knowledge – knowledge about how to actually apply GenAI to solve problems. Whereas declarative knowledge takes the form of awareness and mental representations, procedural knowledge is demonstrated through performance, behavior, and use. Thus, literacy goes beyond mere awareness, although it is often interpreted and operationalized as awareness; in other words, it includes both “knowledge about” and “knowledge how”. It is necessary to further clarify that declarative knowledge is not restricted to metacognitive knowledge, such as acknowledging the importance of AI literacy (e.g., “It's important to learn how to use GenAI”), or confirming knowledge (e.g., “I am aware of the limitations of GenAI”). Measures of GenAI literacy must include items that target specific aspects, such as strategies for effective prompt engineering, as well as activities that require users to demonstrate their GenAI competence through behaviors or task performance.

GenAI literacy is a relatively new concept derived from the notion of AI literacy, so its conceptualization and measurement are similar to general AI literacy. Based on a thorough review of the research, Ng et al. (2021) identified four components for the construct of AI literacy: (1) know and understand, (2) use, (3) evaluate and create, and (4) ethical issues. Ng et al. mapped the first three components onto Bloom's taxonomy

of cognitive skills, which stand in a hierarchy with six levels defined in terms of the information processing demands required at each level. Ranked from the lowest to the highest cognitive demands, these skills include know, understand, apply, analyze, evaluate, and create. Applied to Ng et al.'s (2021) framework of AI literacy, the stage of "know and understand" involves learning the basics of AI including its mechanism and functions; "use" refers to how to apply or use an AI tool to solve problems; and "evaluate and create" are two high-order skills users draw on to assess the quality of AI output and create new applications. The fourth component, ethics, permeates all stages of AI use and does not fall into the hierarchy of cognitive thinking. Drawing on Ng et al.'s framework, Warschauer et al. (2023) proposed a similar framework for GenAI literacy for second language writing, which also applies to L2 learning in general. In this framework, GenAI literacy has five components: understand, access, prompt, corroborate, and incorporate. "Understand" is similar to "know and understand" in Ng et al.'s framework, "access and prompt" corresponds to "use", "corroborate" is equivalent to "evaluate", and "incorporate" has no counterpart in Ng et al., but it can be considered part of ethics as it mainly concerns proper use of GenAI output. Wang and Wang (2025) conducted a small-scale exploratory, observational study investigating 10 L2 English writers' GenAI use and literacy in a writing class at a liberal arts college in the U.S. The L2 writers worked on a writing assignment using ChatGPT. They made reflections on their use of ChatGPT and participated in semi-structured interviews, and the process of their using ChatGPT in writing was captured via a screen recorder. Based on the multimodal data they collected, the researchers proposed a model of critical GenAI literacy, emphasizing the critical dimensions of GenAI literacy. The model consists of four components: critical awareness (akin to "understand" and "know and understand" in the above mentioned models), critical strategies for human-AI interaction (resembling "access and prompt" and "use"), critical evaluation of affordances (similar to "evaluate" and "corroborate"), and critical positionality. Critical positionality, which is missing from other models, refers to maintaining user agency, autonomy, and authorial voice during GenAI use instead of overreliance on it, which may render the author's identity invisible.

In the three models that were discussed above, "know basics of GenAI", "use", and "evaluate" are common components across all the models, although they are labeled slightly differently. Ethics is an independent component in Ng et al.'s model, is included in the "incorporate" component in Warschauer et al.'s model, and is part of the critical awareness component (e.g., bias) in Wang and Wang's model. Critical positionality is a major component of GenAI literacy in Wang and Wang's conceptualization but missing from the other two models. Based on the analysis of the three models and my own understanding, I propose the following components for the construct of GenAI literacy.

**Knowledge of the Basics of GenAI.** The basics of GenAI include, but not limited to, the mechanism of GenAI (e.g., it's based on linguistic instead of semantic probability), historical development, affordances, resources (available tools), controversies, and limitations. This type of knowledge concerns both domain-general information about GenAI and domain-specific information about GenAI's affordances and limitations in L2 learning (including L2 writing).

**Effective Use.** This refers to using GenAI to perform the task or achieve the user's goal, a process that involves what Wang and Wang (2025) referred to as “human-AI interaction”. This component involves knowledge about prompt engineering, such as asking for criteria-based feedback when eliciting corrective feedback on one's writing, and strategies writers utilize in the processes of planning (e.g., outlining), composing, and revising. One insight provided by Wang and Wang's (2025) is that prompt engineering involves human-AI interaction, and principles for effective human communication, such as clarity, specificity, contextualization, etc. are also applicable to our interaction with GenAI.

**Output Evaluation.** This refers to the evaluation of GenAI output to determine if the goal has been achieved. The evaluation may be understood or discussed from two perspectives: criteria for evaluation and methods of evaluation. The criteria for evaluation should be formulated in terms of whether the quality is satisfactory and whether author identity is retained as a result of the appropriation of GenAI output. Quality can be operationalized as accuracy, namely, whether the output, such as feedback generated by GenAI, is accurate, whether the feedback meets the evaluation criteria for the task such as the grading rubric for a writing assignment, or whether the feedback is clear, specific, and supportive – criteria applied in research to evaluate feedback quality. Author identity, which is similar to Wang and Wang's (2025) notion of critical positionality, is essential for the evaluation of the acceptability of GenAI output. The maintaining of author identity can be gauged rhetorically and linguistically, with the former referring to whether the ideational aspects of the author's writing (ideas and organization of ideas) remain after incorporating GenAI output, and the latter to whether the linguistic aspects of GenAI output are in discord with the author's style and identity. For example, in Wang and Wang's (2025) study, one L2 writer refused to accept the sophisticated language suggested by ChatGPT because they felt the language “doesn't sound like me”. Maintaining author agency, autonomy, and identity should occupy a central place in GenAI literacy, because of the possible overreliance on GenAI and the temptation to use GenAI as a replacement, instead of a tool, for human work. Methods of evaluation are related to how to evaluate GenAI output, and one major method is to compare, such as comparing GenAI output with outputs generated by other tools, comparing different

solutions to the same problem suggested by GenAI so as to pick one that meets the user's expectation, etc.

**Ethics.** As discussed below, ethics is a multi-dimensional concept consisting of multiple facets, including bias (underrepresentation of non-English cultures, bias against ESL writers in AI detection, etc.), fairness, breach of intellectual properties in GenAI training, ethical use of GenAI output, and so on. For example, while it is appropriate to incorporate linguistic output from GenAI, it is questionable practice to use GenAI-generated information as research evidence. In Wang and Wang's (2025) study, a student was unable to find sources to support "her argument about Ghanaian family practices" (p. 6). She then cited ChatGPT as "external evidence to back her own experience" (ditto). However, not all information generated by GenAI has an empirical basis, and GenAI output should not be treated as research evidence. Because of the unprecedented ethical concerns GenAI has caused, the importance of ethics in AI literacy is unparalleled. The following section addresses issues surrounding GenAI ethics.

### 3 Ethics

One of the most pressing issues accompanying the advent of GenAI is ethics, which refers to common beliefs of the members of a community or society about the appropriacy of human conduct. In current conceptualization, ethics concerns diversity, equity, and inclusion (DEI), the wellbeing of individuals and the humanity, and social progress. Common principles of ethics may manifest differently within a profession or community. Regarding the use of GenAI, ethics can be understood from the perspectives of *bias*, *ethical use*, *equity*, and *confidentiality*, which are discussed in the following paragraphs.

Bias refers to incomplete or unequal representation of the integral parts, components, members, or perspectives of a unit or community. Bias can take the form of over- or under-representation of certain stake-holders (groups as well as individuals) or entities in a phenomenon, process, or event. GenAI bias takes the form of the over-representation of Western ideology, which underlies the training materials, all of which are written in English and contributed by English speakers. It can also be argued that GenAI outputs are imbued with U.S. ideology and culture because the training materials are in American English; accordingly, other ideologies and cultures are underrepresented. From the standpoint of second language writing and second language learning in general, second language perspectives are missing not

only in training materials, but also in policy-making – L2 users are not considered in policies – as well as access – GenAI outputs are not accessible to L2 users whose English proficiency is often limited. Bias also exists in AI detection, a topic to be revisited below.

Ethical use is probably the greatest concern for GenAI use in writing in general and second language writing in particular. Ethical use concerns whether, when, how, and for what purposes GenAI should and should not be used. These questions should and can be examined empirically. Casal and Kessler (2023) surveyed 27 journal editors about their opinions on acceptable GenAI use for publishing purposes. 16 of them considered it acceptable to use GenAI to edit texts, 14 endorsed using it to generate computer code, 11 felt it Ok to write a summary of an article for public use, and 10 thought it ethical to write the abstract of an article. Few or no editors supported using GenAI to write parts of or the whole main document, and six agreed with the statement that under no circumstance is it acceptable to use GenAI. In general, it would seem that editors considered GenAI a tool that can be used to proofread to improve the clarity and readability of manuscripts, and they thought it unethical to use it to generate new content or perform tasks for which originality is valued.

Evidently, one major topic related to ethical use is plagiarism, which refers to using GenAI-generated output as one's own work without acknowledging the source or using GenAI output in replacement of one's own effort. Plagiarism can be approached and studied from various perspectives, such as plagiarism detection and prevention of plagiarism. Plagiarism detection has received much attention since the launch of ChatGPT. Liang et al. (2023) submitted TOEFL essays written by ESL speakers and essays written by eighth-grade US students to multiple detectors and found that while some detectors correctly recognized eighth-graders' writing as human-written at a high accuracy rate, the detectors misclassified an average of over 60 % of ESL essays as GenAI-generated. Therefore, they concluded that GenAI detectors are biased against ESL writing. They suspected that the bias is due to the mechanism of GenAI detection, which is based on perplexity scores. Perplexity is an index that represents the extent to which language use is predictable, so lower perplexity or higher predictability stands for higher likelihood of GenAI outputs. This method of GenAI detection, which is based on perplexity or predictability, is the same as the mechanism of GenAI per se, which is also based on predictability. Liang et al. (2023) further found that ESL essays' perplexity scores were low, which made the researchers suspect that the high false positive rate for ESL writing was likely due to the low perplexity scores. The low perplexity scores were likely attributable to the lack of linguistic sophistication of ESL writing, which made the writers' language use highly predictable. To verify this hypothesis, they increased the ESL essays' lexical

sophistication by using ChatGPT and found that the inaccurate classification rate was significantly reduced. In another study, conducted by Jiang et al. (2024), a GenAI detector was used to classify human- and GenAI-written GRE essays. The human writers included both native speakers and non-native speakers of English. The researchers found a near-perfect classification rate and did not find any bias against nonnative speakers. There are two explanations for the disparate findings of Jiang et al. (2024) and Liang et al. (2023). One is the higher proficiency of the GRE essay writers in Jiang et al. than the TOEFL essay writers in Liang et al.'s study, and greater closeness to native speakers' writing may have reduced the false positive rate. Another explanation is the training that was conducted in Jiang et al. that enabled the GenAI detector to learn the linguistic features of the genre of GRE writing and give a better detection performance. In this regard, this study suggests the importance of domain-specific training so that GenAI can have sufficient input to extract regularities related to the particular discourse and linguistic features. In general, GenAI detection is still in its infancy because of its high false positive rates, and in fact, based on Dang and Wang's (2024) study on U.S. universities' policies for GenAI use, 61 out of 100 universities officially banned the use of GenAI detection software.

In addition to GenAI detection, Dang and Wang's (2024) also examined university policies on the prevention of plagiarism in writing. Their study showed that the strategies the universities suggested included using personalized assignments such as incorporating students' personal experiences, the local context, etc.; using process-based writing instruction where students are asked to engage in the learning process instead of focusing on the product (e.g., getting a good grade); adopting alternative assignment modalities such as podcasts, presentations, etc. instead of relying on written assignments only; incorporating assignments requiring students to apply skills and knowledge to solve problems. In addition to the above strategies, one way to prevent plagiarism is to require students to disclose whether they used GenAI and for what purposes. However, Tan et al. (2025) found that disclosure of GenAI use in writing led to lower ratings than non-disclosure for the same essays.

Next, I address the last two aspects of GenAI ethics: equity and confidentiality. Equity refers to equal access to GenAI by all individuals regardless of their age, gender, ethnicity, location, nationality, profession, religion, and socioeconomic status. However, universal, equitable access to GenAI is not reality, and many individuals and communities do not have access to GenAI or all features and content because of the lack of infrastructure or because of economic and political reasons. Confidentiality refers to the possible public access to personal information and user responses; however, according to ChatGPT, user responses are not used as training materials, so the concern about confidentiality seems to be assumed.



## 4 Feedback on Writing

Feedback on writing refers to comments on L2 writers' writing performance or quality. In L2 writing research, feedback has been extensively investigated, mainly from the perspective of whether corrective feedback can improve the linguistic accuracy of L2 writing and whether feedback's effectiveness is affected by other factors such as learners' individual differences in anxiety, working memory, etc. (An and Li 2024; Kim and Li 2024; Li and Roshan 2019; Vuogan and Li 2023). Other topics that have been examined include teachers' feedback-providing practices, teacher and student beliefs about corrective feedback, and the congruence and incongruence of teachers' beliefs and their feedback-providing practices (Li 2017). The large amount of experimental and observational research has been synthesized in meta-analyses (e.g., Kang and Han 2015) and narrative reviews (e.g., Li and Vuono 2019). The advent of GenAI marks the beginning of revived and intensified interest in written feedback because providing feedback on writing is a major affordance of GenAI. Besides the replication of traditional research in GenAI contexts, GenAI opens up new research territories, topics, and perspectives, such as comparisons between GenAI and teacher feedback-providing practices, students' engagement with GenAI feedback, etc. GenAI also makes it possible to examine topics that received little attention in previous research such as content- and discourse-related feedback, reformulation of students' essays, etc. In the following sections, I provide a taxonomy of written feedback, synthesize the limited research that has been conducted to date, and discuss the importance and methods of feedback evaluation – a unique topic related to GenAI.

### 4.1 Taxonomy of GenAI Feedback

A taxonomy of feedback is important for a number of reasons. First, users need to know that feedback takes different forms and then use the right prompt to get the feedback they favor. Second, teachers need to be aware that different types of feedback may have differential effects on students' L2 learning and writing and that they should use GenAI to provide the right kind of feedback instead of leaving GenAI to decide what type of feedback is provided. Third, researchers should understand the mechanisms and theoretical bases of different types of feedback and conduct research to empirically verify or examine their effectiveness. Current research simply compares teacher and GenAI feedback as lump sums without distinguishing feedback types, but the findings of the research are not robust and even misleading if teachers and GenAI provide different types of feedback during the instructional

treatment. Some categories of traditional feedback (Ellis 2009; Li and Vuono 2019) apply to GenAI feedback, but feedback also takes new forms in GenAI contexts. Therefore, it is necessary to categorize feedback types so as to have a clear idea of what can be examined in further research. As displayed in Table 1, in general, feedback can be divided into two large categories based on the target and characteristics of feedback. Under the target of feedback, a distinction can be made between global and local feedback according to the aspects of writing that receive feedback. Global feedback targets aspects that influence the overall quality of writing, such as content, organization (flow of ideas, move structure, etc.), and meta-discourse (engaging with audience), etc.; local feedback concerns language and mechanics. Depending on the scope of the target, feedback is either focused or unfocused; the former focuses on particular aspect of writing such as the topic sentence or the English passive voice while the latter has no focus or targets all aspects of writing.

Along the dimension of properties of feedback, feedback can be further classified based on the action taken on the target, the context in which feedback is provided, the language of feedback, and the function of feedback. Action taken refers to what is done on the error or the targeted aspect, which yields the following subcategories:

- *Direct correction*: replacing the target with the correct or another form
- *Metalinguistic feedback*: providing comments on the nature of the flawed target and/or clues on how to improve;
- *Indirect feedback*: highlighting the erroneous part without providing the correct form
- *Reformulation*: rewriting a sentence or a bigger unit such as the whole text without altering the meaning
- *Modelling*: providing a model text on the same topic or based on the same prompt without taking any action on the student's writing.

Reformulation and modelling are not commonly examined in existing feedback research. Reformulation can be partial or comprehensive, depending on the level of changes made to the writer's written output. In the literature, however, reformulation is often defined as correcting errors without marking the corrections, which can be understood as direct correction without tracked changes (e.g., Kim and Bowles 2019). Regarding modelling (Nguyen et al. 2024), one may argue that it cannot be regarded as feedback as it does not include comments on students' writing. However, the counterargument is that it is a response to students' imperfect writing with the intention to improve their writing by encouraging them to notice the gap between the correct model and their own output; therefore, it serves the same function as feedback. It is important to clarify that the aforementioned feedback categories apply primarily to linguistic errors, and to date, feedback on the content and

**Table 1:** Taxonomy of written feedback.

Target of feedback	Aspect of writing	Global	Content, organization, meta-discourse, clarity, readability
		Local	Grammar, vocabulary, mechanics
	Scope	Focused	Targeting a particular linguistic structure such as the past tense, or a particular aspect of writing, such as the topic sentence
		Unfocused	Targeting all aspects
Properties of feedback	Action taken	Direct correction	Replacing an error with the correct form
		Metalinguistic feedback	Commenting on the nature of an error or aspect of writing
		Indirect feedback	Highlighting errors
		Reformulation	Changing the erroneous (and other) parts of a sentence or a bigger unit without changing the meaning and marking the changes
	Context	Modelling	Providing a model text
		Integrated	Feedback is embedded in the text
		Detached	Feedback is separate from the text, such as a list of corrections or reformulated sentences
	Language	Interspersed	Feedback is provided in blocks alternating with text
		L1	Feedback is provided in the writer's native language
		L2	Feedback is provided in the writer's second language
Source	Function	Corrective	Correcting errors or making improvements
		Confirmative	Positive comments; confirming strengths
	Configurations	Teacher	Feedback is provided by the teacher
		GenAI	Feedback is provided by GenAI
		Peer	Feedback is provided by peers
	Configurations	Teacher-adapted GenAI feedback; teacher feedback + GenAI feedback; GenAI feedback + peer feedback, teacher feedback + peer feedback + GenAI feedback, etc.	

organizational aspects of writing has received little attention. In existing feedback research, such feedback is simply labelled as a focus or target (e.g., “content-related feedback”), and further classification in terms of specific focus and actions taken awaits further research.

The context in which feedback is provided differentiates integrated, detached, and interspersed feedback. Integrated feedback is embedded within the text, interspersed feedback in provided in blocks alternating with the text, and detached

feedback is separate from the main text such as in the form of a list of corrections or suggestions. The language of feedback can be the user's first or second language, which bears on the accessibility of feedback to second language users. The final dimension, function of feedback, concerns whether feedback aims to improve or confirm L2 writers' output. The source of feedback refers to the provider of feedback, and there are essentially three sources of feedback: teacher, GenAI, and peers. The three sources can generate different feedback configurations or packages, such as teacher-adapted feedback where the teacher makes changes to GenAI feedback; teacher + GenAI feedback where the teacher and GenAI focus on different aspects of students' writing; GenAI + peer feedback where GenAI and peers give feedback on other students' writing or engage in other activities such as working together to evaluate the quality of GenAI feedback.

## 4.2 Feedback Evaluation

Feedback evaluation refers to efforts to evaluate the quality and acceptability of GenAI feedback for research or pedagogical purposes. The evaluation of feedback quality is a necessary step and a unique topic related to the use of GenAI. It is necessary because GenAI feedback is often inaccurate and inconsistent. For example, Koltovskaia et al. (2024) found that half of the feedback provided on graduate students' academic writing was inaccurate. Lin and Crosthwaite (2024) showed that ChatGPT gave different kinds of feedback on different essays even though the prompt or instructions for feedback elicitation were the same. The consequence of inaccurate feedback is obvious: it is misleading and may have an unfavorable impact on writing quality. Therefore, researchers, teachers, and learners/users should take a critical stance on GenAI feedback, appraising its quality instead of accepting it without scrutiny. In addition to ensuring feedback quality, another function of feedback evaluation is pedagogical: by evaluating GenAI feedback, students reflect on their own written output and compare it with GenAI output, which may facilitate L2 development, learner autonomy, and critical thinking skills.

The evaluation of feedback quality should be based on criteria, and what constitutes valid criteria is a theoretical and empirical question that needs to be examined in research. Lin and Crosthwaite (2024) classified feedback into three categories: accurate, inaccurate, and redundant, with redundant feedback referring to feedback that is unnecessary, although it is not wrong. Steiss et al. (2024) evaluated ChatGPT and human feedback based on whether (1) it is based on criteria, (2) the instructions are clear, (3) it is accurate, (4) essential features are prioritized, and (5) a supportive tone is used. The five criteria are further discussed below.

To begin with, criteria-based feedback refers to feedback provided with reference to certain benchmarks, expectations, or standards. Criteria are of two types: curriculum-based and assessment-based, with the former referring to the expected learning outcomes of a course or training program, and the latter to assessment criteria for large-scale tests such as TOEFL, IELTS, or the Band-4 English test for Chinese university ESL learners. Criteria-based feedback is important because GenAI feedback has been found to be inconsistent and unsystematic (Lin and Crosthwaite 2024), partly because there is a lack of criteria. Thus, GenAI's feedback-providing behaviors need to be regulated by criteria, which serve as user's instructions to GenAI on what type of feedback it is expected to provide and what expectations should be met. The need for criteria is also because what constitutes effective writing depends on the nature of the writing, genre, purpose, audience, task, etc. There are no universal criteria for high-quality writing. What is considered appropriate for an email may not be appropriate for argumentative writing. Likewise, criteria for effective narrative writing are different from criteria for expository writing. For example, a prompt that says "If you were to like grade this, what would be my grade? How would you suggest to make this better?" (Wang and Wang 2025, p. 6) will not generate constructive or satisfactory feedback that improves the essay's grade because GenAI's was not informed of the criteria to evaluate and improve the quality. The second criterion for good feedback concerns clarity, that is, good feedback is clear and unambiguous and provides straightforward information on what action should be taken. The third criterion concerns accuracy, which is also a criterion used by Lin and Crosthwaite (2024). Accuracy is not restricted to language, such as advising the author to add a plural -s to the word damage in the phrase "a lot of damage" (Lee 2004); it also targets feedback on other aspects of writing, such as suggesting a topic sentence that is unrelated to the paragraph in question. The fourth criterion, prioritizing essential features, is not straightforward. Based on Steiss et al.'s (2024) explanation, it has to do with (1) whether the feedback focuses on the most important aspects of writing given that too much feedback is overwhelming, and (2) whether the feedback suggests an action that is feasible or within the writer's capability. The former is related to amount and the latter to practicality. The fifth criterion refers to whether the feedback consists of positive remarks such as a praise or facilitative language (e.g., "Adding more evidence will strengthen your argument").

Some clarification is in order. First, it is important to point out that what constitutes good feedback is an empirical question. There are various ways to formulate and validate feedback evaluation criteria, such as asking stakeholders (e.g., teachers and students) for their views and perceptions, asking experts to identify good and poor feedback, or asking teachers to adapt feedback and then extract themes based on their adaptations. Second, feedback evaluation criteria should be domain-specific,

for example, one criterion for feedback evaluation in L2 writing is the accessibility of feedback, that is, whether L2 writers can easily understand the language and content of the feedback. Feedback accessibility is important because L2 writers may not be able to understand the feedback if it is delivered using sophisticated vocabulary and linguistic structures. Accessibility is more relevant to L2 writing than L1 writing where native speaking writers normally do not have difficulty with the language of feedback if it is delivered in their first language. Third, feedback evaluation should be restricted to the quality and attributes of feedback per se and should be teased out from the choice of feedback, namely which types of feedback should be provided, because decisions on what feedback should be provided must be based on experimental research that can show which feedback type is more effective than other feedback types in improving writing quality or facilitating learning gains.

### 4.3 Research Findings

In general, empirical research on GenAI feedback can be divided into two broad categories: observational and interventional. Observational research is exploratory and descriptive and examines GenAI's feedback-providing practices (normally in comparison with teachers' feedback) and teachers' and students' attitudes toward GenAI feedback, students' engagement with GenAI feedback, etc. Interventional research is experimental, examines causal effects, and focuses on the effectiveness of GenAI feedback in facilitating L2 writing development. Most existing research is observational, which is not surprising given that GenAI is a recent innovation. The following sections provide a synthesis and critique of the findings of the two streams of research, and due to limited research, the findings are suggestive rather than conclusive.

Table 2 shows the findings of observational research. One recurring theme that emerged from existing research is that GenAI excels in providing organization-related feedback, while teachers are better at providing content-related feedback (Guo and Wang 2024; Zou et al. 2025). The studies show that compared with teachers, GenAI provided more feedback on organization, their feedback on organization was more actionable, students' revisions were more successful after receiving organization-related feedback from GenAI, and students were also more positive about GenAI-generated feedback on organization than teachers' feedback on organization. Compared with GenAI, teachers provided more feedback on content and students were more successful in their revisions in response to teacher feedback on content. Regarding the types of feedback favored by GenAI and teachers, the picture is less clear, but GenAI tended to provide more direct feedback (informing learners what to do, providing the correct form, and reformulating the original text), while

teachers provided more indirect feedback (Guo and Wang 2024; Lin and Crosthwaite 2024). In terms of the amount of feedback, GenAI provided more feedback than teachers in an unmonitored setting (Guo and Wang 2024), but they provided similar amounts of feedback when teachers were encouraged or instructed to provide feedback (Zou et al. 2025). Han and Li (2024) reported that students received an average of 13.38 cases of teacher-adapted GenAI feedback for the first essay and 7 cases for the second essay. The ideal amount of feedback is an empirical question that has yet to be examined in research, but research on traditional feedback shows that “less is more”, which means that a smaller amount of feedback is likely more effective than excessive feedback (Kang and Han 2015; Lee 2008; Mao et al. 2024).

Students' engagement with feedback is a tripartite construct defined as students' affective, behavioral, and cognitive involvement with feedback. Affective engagement concerns users' attitudes and emotional reactions to feedback. Zou et al. (2025) reported that students were more positive about teacher feedback than GenAI feedback, demonstrating students' greater confidence in teacher feedback. However, Zou et al. further showed that there was an interaction between the source and target of feedback. Specifically, students were more positive about GenAI-generated than teacher-generated organization-related feedback, but they were more positive about teacher-generated content-related feedback than GenAI-generated content-related feedback. One caveat is that in this study, teacher and GenAI feedback was provided together in the same text and students were informed of the different sources of feedback, which may have impacted their reactions and attitudes. Escalante et al. (2023) found that students were divided between favoring GenAI and teacher feedback. Behavioral engagement is represented by what students do in response to feedback such as students' incorporation of feedback in revisions. According to Zou et al. (2025), students made more revisions in response to teacher feedback than GenAI feedback, although similar to what was found about affective engagement, students' success rate was higher for GenAI feedback on the organizational aspects of writing than teacher's feedback on organization. Koltovskaia et al. (2024) and Han and Li (2024) showed that L2 writers incorporated or accepted over 60 % of GenAI feedback. Cognitive engagement refers to mental or psychological behaviors that are not visible such as efforts to understand, analyze, and evaluate feedback. Koltovskaia et al. (2024) revealed a high level of cognitive engagement in GenAI feedback in L2 writers who noticed errors and spotted inaccurate feedback.

The quality of feedback also figured in the research. Feedback quality was evaluated by means of teachers' and students' perceptions and content analysis of feedback. In general, GenAI feedback was considered more specific and accurate than teacher feedback and more versatile than automated writing evaluation

Table 2: Findings of observational research.

Dimensions		Findings	
		Suggestive trends	Evidence
Type of feedback	Target of feedback	GenAI feedback focuses more on organization while teacher feedback more on content	GenAI provided more actionable feedback on <u>organization</u> while teachers provided more actionable feedback on <u>content</u> (Zou et al. 2025) GenAI feedback focused more on <u>organization</u> while teacher feedback focused more on <u>content</u> ; both provided lots of feedback on <u>language</u> (Guo and Wang 2024)
	Action taken on target	GenAI provides more direct feedback while teachers provide more indirect feedback	GenAI provided more <u>directive</u> feedback (telling students what to do); teachers provided more <u>informing</u> (making comments) and <u>query</u> feedback (asking for clarification–indirect) (Guo and Wang 2024) GenAI provided more <u>reformulation</u> + meta-linguistic feedback; teachers provided more direct correction and <u>indirect</u> feedback (error codes, queries, comments) (Lin and Crosthwaite 2024)
Amount of feedback		GenAI provides more feedback than teachers in the natural state	GenAI provided more feedback than teachers in an unmonitored condition (Guo and Wang 2024) GenAI provided similar amounts of feedback as teachers when teachers were encouraged or instructed to provide feedback (Zou et al. 2025) Students received an average of 12.38 cases ChatGPT-supported feedback (teacher-adapted ChatGPT feedback) for Essay 1, and 7 such cases for Essay 2 (Han and Li 2024)
Students' engagement with feedback		Affective engagement: students' attitudes toward GenAI feedback vis-à-vis teacher feedback	Students were more positive about teacher feedback than GenAI feedback when both types of feedback were provided (Zou et al. 2025) Students were more positive about organization-related feedback from GenAI than from teachers; however, they were more positive about content feedback from teachers than GenAI (Zou et al. 2025) Students were split between favoring GenAI and teacher feedback via conferencing (Escalante et al. 2023)



Table 2: (continued)

Dimensions		Findings	
		Suggestive trends	Evidence
Quality of feedback	Pros	Behavioral engagement: revisions after receiving feedback	Students' integration of teacher feedback and the success rate was higher after receiving teacher feedback than GenAI feedback (Zou et al. 2025) Students engaged more with GenAI's feedback on organization but more with teachers' feedback on content and language (Zou et al. 2025) Students received 134 comments from GenAI, accepted 87, and dismissed 47 (Koltovskaia et al. 2024) 64.54 % and 63.91 % of the changes in Essay 1 and Essay 2 were successful revisions (Han and Li 2024)
		Cognitive engagement: mental effort and processing	All learners noticed errors and spotted inaccurate feedback (Koltovskaia et al. 2024)
		Specific	GenAI feedback was more <u>specific</u> and accurate than teacher feedback (students' perceptions) (Escalante et al. 2023) GenAI feedback was detailed, <u>specific</u> , and flexible compared with automated writing evaluation tools such as Grammarly (teachers' perceptions) (Guo and Wang 2024) GenAI feedback was comprehensive, accurate and <u>specific</u> (students' perceptions) (Zou et al. 2025)
	Cons	Versatile	GenAI feedback addressed <u>content and organization</u> while automated writing evaluation systems such as Grammarly mainly focused on <u>language</u> (teachers' perceptions) (Guo and Wang 2024)
		Accurate	When direct correction was provided, GenAI was more <u>accurate</u> than teachers (content analysis) (Lin and Crosthwaite 2024) GenAI feedback was comprehensive, <u>accurate</u> and specific (students' perceptions) (Zou et al. 2025)
		Inaccessible	GenAI feedback was lengthy and inaccessible (teachers' perceptions) (Guo and Wang 2024) GenAI feedback was unnecessarily sophisticated (students' perceptions) Koltovskaia et al. (2024)

Table 2: (continued)

Dimensions	Findings	
	Suggestive trends	Evidence
	Not individualized	GenAI's feedback criteria are different from teachers'; it's not tailored to students' back-ground and the local curriculum (teachers' perceptions) (Guo and Wang 2024) GenAI feedback may cause students to lose identity (students' perceptions) Escalante et al. (2023) Students preferred teacher feedback over GenAI feedback because the former was personalized (Zou et al. 2025)
	Inaccurate and redundant	Half of GenAI feedback was <u>inaccurate</u> (Kol-tovskaia et al. 2024) GenAI suggested <u>unnecessary</u> , albeit not incorrect, changes (Lin and Crosthwaite 2024) GenAI failed to recognize <u>irrelevant</u> content (teachers' perceptions) (Guo and Wang 2024)
	Confusing display	GenAI used sentence numbers instead of quoting the wrong sentences (Koltovskaia et al. 2024) GenAI feedback was difficult to locate (teach-ers' perceptions) (Guo and Wang 2024)
	Inconsistent	GenAI feedback took varied forms across different texts even though it was based on the same prompt (Lin and Crosthwaite 2024)

systems such as Grammarly, whose feedback primarily focuses on language and mechanics (Escalante et al. 2023; Guo and Wang 2024; Zou et al. 2025). However, GenAI feedback has been found to have a number of limitations. The studies showed that GenAI feedback is inaccessible and sophisticated, posing a challenge for L2 learners; it lacks individualization, namely it treats all writers in the same way and ignores their background, experience, L2 proficiency, and the local curriculum; it can be inaccurate and redundant, with redundancy meaning that the feedback is unnecessary although it is not inaccurate; its format and display can be confusing, for example, it uses sentence numbers rather than quotes the wrong sentences; the feedback type is inconsistent across essays even if it is given the same prompt (Escalante et al. 2023; Guo and Wang 2024; Koltovskaia et al. 2024; Lin and Crosthwaite 2024).

The research discussed above is observational, and there is limited intervention research examining the effects of GenAI feedback on L2 writing development. In this section, I review two quasi-experimental studies comparing the effects of GenAI and teacher feedback. Boudouaia et al. (2024) conducted a study involving 76 Algerian university EFL learners, who were divided into experimental and control groups and completed four argumentative writing tasks. The experimental group received feedback from ChatGPT and revised each essay, while the control group received teacher-led writing instruction, the details of which were not provided. The results showed improvements in students' rated writing and their acceptance of ChatGPT. The ChatGPT group outperformed the teacher-led group. In another study, Escalante et al. (2023) examined the effects of GenAI feedback and feedback provided by a tutor during weekly conferencing sessions on 48 university EFL students' writing improvement. The students were divided into two groups and received feedback on their writing for six weeks. GenAI feedback targeted both language and content, and language feedback included a list of errors, descriptions of the nature of errors, and suggested corrections. The content of teacher feedback was unclear. The study found no difference between GenAI feedback and teacher feedback. In these two studies, the type and amount of feedback were not controlled for, making it impossible to attribute treatment effects, if any, to the manipulated variable, which is GenAI feedback versus teacher feedback.

## 5 Prompt Engineering

A prompt refers to instructions given to GenAI to perform a task or fulfil user requirements. Prompt engineering is the process of optimizing the instructions for GenAI to achieve a desired outcome. Prompt engineering is a unique topic for GenAI, and it is a crucial step because the quality of GenAI output is determined by the prompt it receives. A good prompt requires a certain level of GenAI literacy such as its affordances and principles of prompt engineering as well as domain-specific knowledge about the task, topic, and discipline. A good prompt should be clear, specific, and contextualized. Clarity is achievable by using unambiguous language, avoiding expressions with multiple interpretations, clarifying concepts that are not straightforward, and exemplifying expected output. Specificity requires the user to provide all the details necessary for the successful performance of the task GenAI is entrusted with, including input materials, action to be taken by GenAI, and output specifications. Contextualization pertains to providing information about the setting for the task GenAI is asked to perform, stakeholders' characteristics (such as the writer's L2 proficiency, age, grade level), and the audience or consumers of GenAI output.

An ideal prompt for GenAI use in L2 writing should consist of the following three components, illustrated by using the example of written feedback.

- *Input*: materials GenAI works with. In the case of feedback on writing, input materials include (1) the writing task, including the exact writing prompt and its requirements such as time limit and word limit as well as the nature and goal of the writing task, and (2) the writers, such as students' background (e.g., 8th-grade ESL learners at a Korean middle school), age, and English proficiency.
- *Task*: the duty GenAI is asked to perform. This component also entails assigning a role for GenAI, for example, "You are an ESL teacher, and your job is to give feedback on students' writing".
- *Output*: the outcome to be generated by GenAI. If GenAI is asked to provide feedback, this component should include information about the type of feedback (global or local; direct, indirect, or metalinguistic; focused or unfocused); the visual display of feedback (integrated, interspersed, or detached); the language (L1 or L2) in which feedback is given; the audience or consumer of the feedback (e.g., ESL writers); the criteria or rubric for evaluating students' writing performance so that GenAI feedback is aligned with curricular goals.

Several strategies for prompt engineering should be highlighted. One is training, where GenAI is provided with a model performance or some vicarious experience that enables it to learn from examples. Prompts that involve exemplification and demonstration are called few-shot prompts, which are distinguished from zero-shot prompts where no examples or demonstrations are given. Of course, zero-versus few-shot prompts should not be taken as a dichotomy but a continuum instead. Complex tasks such as giving feedback on an essay or evaluating writing performance require few-shot prompts. Simple tasks can be completed with zero-shot prompts, such as "Please correct the grammar in the following paragraph" (Wang and Wang 2025, p. 6). Another strategy is to divide a complex task into several iterations, each focusing on one subtask, such as targeting different aspects of writing in the case of written feedback.

To conclude this section, I present the prompts used in some studies on L2 written feedback to exemplify the variation of prompts across studies, which may have potentially caused disparate results regarding GenAI's feedback providing practices. In Guo and Wang (2024), GenAI was asked to provide feedback on content, organization, and language separately in three rounds, which allowed the researchers to explore the different foci of GenAI and teacher feedback. However, the teacher participants did not seem to be required to provide feedback on the three aspects in three rounds. It is unclear how this minor variation may have influenced the results of the study. Koltovskaia et al. (2024) provided example prompts to students for use in different stages of writing such as brainstorming, outlining, drafting,

etc. To receive feedback, GenAI was asked to give “some constructive criticism” (p. 1, Supplementary Information), identify grammar and punctuation errors and suggest corrections, etc. Lin and Crosthwaite (2024) prompted GenAI to provide feedback on grammar, vocabulary, organization, and ideas, but GenAI was given the freedom to provide feedback in “any form” and “use one or numerous strategies” (p. 6). The prompt of Zou et al.’s (2025) study specified content, organization, and language as the aspects for GenAI feedback and required feedback to be given in Chinese, the students’ L1. As can be seen, these prompts have commonalities and disparities, but slight disparities may cause a major impact on GenAI’s feedback-providing practice.

## 6 Writing Assessment

GenAI can be used to assess writing quality or grade students’ writing assignments. To be clear, assessing is different from giving feedback because the former aims primarily to evaluate writing quality and the latter seeks to enhance writing quality and facilitate learning. A major topic for research is whether GenAI assessments are valid and reliable, and the predominant method used in the research is to compare GenAI and human ratings on the same essays following the same rubric to see whether the two ratings are consistent or correlated. A strong correlation suggests that GenAI ratings are comparable with human ratings and are therefore valid. Cast in assessment terms, this type of evidence is collected to verify a test’s concurrent validity, namely whether the test is comparable with an alternative test that measures the same construct, especially a test that has proven valid. It must be clarified that concurrent validity rests on the assumption that the criterion or existing test is valid; thus, if the test being validated is not correlated with the existing test, it means the test under validation is invalid. However, in the case of GenAI assessment, we cannot conclude that GenAI assessment is not valid if it is not aligned with teacher assessment, unless the validity of teacher assessment is established. In addition to current validity, the validity of a test can be verified from other perspectives, such as predictive validity, which concerns whether scores of a test can predict an outcome, or divergent validity, which refers to whether a test is uncorrelated with a test that is hypothesized to be unrelated to the test in point.

Among the four studies discussed below (Table 3), all examined concurrent validity by comparing GenAI ratings with human ratings or an automated writing evaluation system such as Grammarly; one examined predictive validity by probing whether GenAI ratings for linguistic accuracy can predict students’ overall writing quality. The studies to be reviewed fall into two categories investigating GenAI’s validity in assessing *overall* writing quality (Bucol and Sangkawong 2024; Shin and Lee 2024) and specific aspects of writing such as accuracy, complexity, and fluency

(Mizumoto et al. 2024; Yamashita 2024). Bucol and Sangkawong (2024) sought to validate GenAI's ratings for a descriptive writing task completed by EFL learners at a university in Thailand. 10 teachers were asked to rate 10 students' essays; five teachers used ChatGPT and five used their own judgments based on the same criteria, which were related to task achievement, grammar, lexical selection, logic, and mechanics. The correlation between the two types of ratings was  $r = 0.65$ ,  $p = 0.04$ . Despite the authors' conclusion that ChatGPT demonstrated significant benefits such as efficiency, consistency, etc., a correlation of 0.65 shows a certain degree of inconsistency between human and ChatGPT ratings and is not satisfactory for assessment purposes. The authors presented snapshots of human-ChatGPT exchanges exemplifying ChatGPT's failure to consult certain criteria and its omission of errors. The article did not report any prompt engineering where ChatGPT and human ratings would have been calibrated to be aligned with each other. It is also unclear if criteria that are not straightforward such as "task achievement" and "logic" were defined and exemplified. Perhaps prompt engineering and training would have led to greater consistency. These speculations were confirmed by another study, conducted by Shin and Lee (2024).

In Shin and Lee's study, 50 persuasive essays (80–120 words) written by Year 2 students from several Korean high schools were rated by two experienced high school teachers who were certified writing examiners. The same essays were also rated by a scoring system built using ChatGPT. Both human and GenAI (ChatGPT) raters were provided with detailed guidance on what to do in the rating process, a detailed scoring rubric, sample essays, and sample scores. The authors reported piloting the scoring tool to ensure it could score essays according to instructions. Although details about the piloting process were not provided, clearly some training may have occurred to improve the tool's scoring performance. The results showed high consistency between the two raters and ChatGPT, with correlations ranging between 0.78 and 0.87 and only one correlation falling below 0.80. The higher human-GenAI consistency seems to be primarily due to the rigorous training GenAI received that included not only the scoring rubric but also sample scored essays; the accurate ratings provided by the two human raters; the clearly defined rating criteria; and the larger score range (1–5) than Bucol and Sangawong (2024) (0.5–2).

While the two above studies examined GenAI's ability to evaluate overall writing quality, Mizumoto et al. (2024) and Yamashita explored GenAI's effectiveness in rating specific aspects of L2 writing. Mizumoto et al.'s study focused on accuracy and found an overall correlation of 0.79 between GenAI and human ratings of ESL writing accuracy; the correlation was stronger than the correlation between the ratings of GenAI and Grammarly,  $r = 0.69$ . GenAI's correlation with overall writing was also stronger than Grammarly's correlation with overall writing: 0.63 versus 0.55. Although a correlation of 0.79 between GenAI and human raters is acceptable, it is

**Table 3:** A comparison of studies on the validity of GenAI-based writing assessments.

	<b>Bucol and Sanga-wong (2024)</b>	<b>Shin and Lee (2024)</b>	<b>Mizumoto et al. (2024)</b>	<b>Yamashita (2024)</b>
Rated aspects	Overall quality: task achievement, grammar, lexicon, logic, mechanics	Overall quality: task completion, content, organization, language	Accuracy	Complexity, accuracy, and fluency
Rating method	Judgement	Judgment	Calculation of error rate	Judgment
Feedback	ChatGPT was asked to give feedback, score, and suggestions for improvement	ChatGPT was asked to provide feedback and a revised essay	GenAI must provide corrected sentences	No feedback required
Human raters	10 Thai university EFL teachers	2 certified high school teachers	Not reported	80 native and non-native speakers of English
Essays	10 90- to 250-word descriptive essays	50 80- to 120-word persuasive essays	232 letter-type 120–180 word essays from the Cambridge English test	140 argumentative 200–300 word essays from the ICNALE corpus
Writers	University EFL students	Second-year high school students	Asian EFL learners	136 nonnative and 4 native speakers
Prompt engineering	Zero-shot; prompt included a rubric and the role for GenAI (university EFL teacher)	Few-shot; prompt included rubric, sample essays, writer identity, training	Zero-shot; GenAI was told to be strict	Zero-shot; prompt included information on complexity, accuracy, and fluency; GenAI told to be lenient and not to provide any justification
Score range	0.5–2.0	1.0–5.0	Errors per hundred words	0–10
GenAI-human consistency ( <i>r</i> )	0.65	0.87, 0.88, 0.89, 0.87, 0.78, 0.80, 0.81, 0.82	0.79	0.67, 0.82, 0.75

below 0.80, which the authors stated may have been due to the disparity in the interpretation of what constitutes an error. For example, while human raters ignored punctuation, GenAI consistently counted punctuation-related errors. The

researchers also acknowledged that the inconsistency was perhaps partly because they used a zero-shot prompt that lacked elaboration and exemplification of the rating criteria. Yamashita (2024) compared GenAI and human evaluations of complexity, accuracy and fluency, and found that the GenAI-human rating correlation was 0.67 for complexity, 0.82 for accuracy, and 0.75 for fluency. Overall, these correlations are relatively weak, albeit acceptable. There are several caveats about Yamashita's prompt. First, explanations on the three rated dimensions are not straightforward. For example, complexity is defined as the extent to which "you think the speaker/write uses morphologically and/or complex words, phrases, expressions, constructions, and grammar" (p. 6). But it is unclear what specifically complex words and grammar mean. In the literature (e.g., Li and Fu 2018), syntactic complexity has been operationalized in various ways, such as subordination, average sentence length, etc., and so has lexical complexity, which can be indexed by lexical variety (using different words), lexical sophistication (using less frequent words), etc. Also, in this study, GenAI was informed that the essays were written by nonnative speakers of English and should be graded accordingly. However, GenAI was not provided with any further information about the norms for nonnative speakers. Fluency has been operationalized as the number of words a text contains in writing research. But in Yamashita's study, fluency was primarily defined as dysfluencies, such as pauses, hesitations, false starts, etc., information that is not relevant to the rated texts.

## 7 Discoursal Comparison Between GenAI and Human Writing

One promising area of research is the comparison between GenAI- and human-generated texts in terms of linguistic, rhetorical, and semantic features. In the literature, there are essentially two types of methods to examine this topic: direct and indirect. In direct methods, human writers and GenAI are asked to write essays based on the same prompt, and a corpus analysis is conducted to identify the differences and commonalities between the two sets of essays (Jiang and Hyland 2024). One principle for this method is to use the exact same writing prompt for both human writers and GenAI in order to ensure the two types of written output are comparable. It is also necessary to specify the identity GenAI assumes (e.g., "You are a second-year student at the Department of English of X University"). In indirect methods, differences between GenAI and human writing are *inferred* based on GenAI's feedback on



human writing (Wang and Wang 2025) or its adaptations to human-generated texts. In the case of adaptations, two approaches have emerged: one is to ask only GenAI to adapt (Huang and Deng 2025), and the other is to ask both GenAI and human editors to adapt the same ESL essays and compare their adaptations (Tu 2025). A comparison of human and GenAI writing is beneficial in at least two ways. The results may (1) facilitate GenAI plagiarism detection by helping the tool recognize features of GenAI writing, and (2) raise writers' awareness of the features of GenAI writing and help them avoid those features to distinguish their own writing from GenAI writing. Avoiding features of GenAI writing is particularly important for L2 writers because L2 writing is more easily misclassified as GenAI-generated compared with texts authored by native speakers, as previously discussed. The following sections provide a snapshot of the findings of the research comparing GenAI and human writing. Although not all studies involve L2 writers, the findings have implications for L2 writing.

**High Predictability.** Jiang and Hyland (2024) found that student writers used more clause-related bundles, such as *that they are* and *as it is*, which involve complicated relationships between linguistic units in a sentence. GenAI writing, however, contained fewer clause-related bundles, which are less predictable. Jiang and Hyland argued that this finding suggests that human writers are more reliant on context instead of the local, isolated collocations of words – formulaic linguistic sequences serving as the bases of GenAI's input. The study also shows that GenAI essays contained more noun-related chunks than human writers, such as *the development of*, which, according to the authors, is due to the greater predictability of noun phrases than other bundle types.

**Lack of Authorial Presence/Voice.** According to Jiang and Hyland (2024), student writing contained more human subjects, such as *I believe that*, *some argue that*, and explicit self-mentions, such as *in my opinion*, while human subjects are missing in GenAI writing, demonstrating more authorial presence in human writing. Echoing Jiang and Hyland's finding, Huang and Deng (2025) reported that GenAI replaced first-person pronouns with shell nouns when asked to adapt dissertation abstracts. For example, *I present evidence for black hole spin...* was replaced by *The analyses yield evidence supporting the presence of black hole spin*, which made the author invisible.

**Focus on Linguistic Cohesion.** According to Jiang and Hyland (2024), GenAI used more organizing devices such as (on) *the other hand*; *not only... but also...*; *in conclusion*. Human writing contained more causal or resultative signals, such as *as a result*, *due to the*, and framing devices such as *if there is*, *in a way*, etc., reflecting

human beings' logical thinking and reasoning ability, which stands in contrast with GenAI's surface cohesion between the different parts of the local text. For example, Tu (2025) discovered a striking change GenAI made to ESL essays: adding more coordinating conjunctions (*and, or*).

**Academic and Technical Themes.** Zhang and Crosthwaite (2025) compared GenAI and L2 human writing and found that while GenAI essays addressed academic and technical themes, L2 writers focused on personal and social issues. The researchers argue that the differences in the themes addressed are reflections of linguistic competence. That is, GenAI has a more varied, larger lexical repertoire, which allows it to generate texts on academic and technical themes. L2 writers' limited linguistic knowledge is sufficient only for familiar and personal themes. However, an alternative interpretation is that GenAI writing focuses on topics that are impersonal, formal, and academic, whereas human writers are more interested in social phenomena, personal well-being, and humanitarian issues. It would be interesting to compare the themes addressed in GenAI and human writing by native speakers who do not have language deficits or disadvantage.

**Linguistic Sophistication.** In Wang and Wang (2025), students of a writing class at a liberal arts college in the U.S. used ChatGPT to complete a writing assignment and reflected on their experience. Some students rejected ChatGPT's feedback on their language use because the suggested language "sounded unnecessarily complicated" (p. 12). One student commented that sophisticated words such as *penchant* and *nuances* normally do not appear in her own writing.

## 8 Implications, Future Directions, and Guiding Principles

In the final section of the article, I discuss the implications of the research, tentative conclusions that can be drawn, future directions for major strands of research, and underlying principles for GenAI use based on the findings of L2 writing research. Starting with GenAI literacy, it is crucial to understand GenAI's mechanisms, affordances, and limitations. GenAI literacy is an important research topic that needs to be conceptualized, instruments must be developed to measure it, and research must be conducted to examine the effects of interventions on its improvement. Due to the importance of GenAI literacy and users' lack of knowledge about its affordances and limitations, GenAI literacy should be included as part of GenAI interventions L2

writing research (Koltovskaia et al. 2024). It should also be a pedagogical requirement that must be addressed in institutional policies, course syllabi, and program curricula where independent courses can be offered on GenAI literacy. One aspect of GenAI literacy is ethics related to GenAI use, and two core components of the GenAI ethics are bias and ethical use. Biases might be political, ideological, cultural, and social, and one bias related to L2 writing is the bias against ESL writers – they are often misclassified as GenAI writers. Ethical use is a central topic of GenAI use because one of the greatest challenges posed by GenAI is preventing plagiarism and fraud. Both bias and ethical use require empirical investigation that can be conducted from various perspectives.

Moving on to the process of using GenAI, one crucial initial step is prompt engineering because the quality of GenAI relies on the prompt. A good prompt must be clear, specific, and contextualized. It is important to specify all the parameters of the expected output, define and exemplify concepts, provide enough “fodder” – training materials – to GenAI so it can recognize patterns and learn what it is expected to do. When eliciting feedback on writing, the user needs to give specific instructions on what types of feedback are expected, provide criteria for the expected feedback, and vet the quality of the generated feedback. The research on corrective feedback suggests that teachers and GenAI may work together and draw on their strengths in feedback-providing, for example, GenAI may focus on organization and teachers on content. Peer feedback may be drawn into the picture and may focus on content, together with teacher feedback, given the effects of content-related peer feedback on L2 writing (Vuogan and Li 2023). It must be pointed out that most research on GenAI feedback is observational, examining learners’ attitudes, use, and engagement with GenAI feedback, and there is insufficient experimental research on GenAI’s effects on the development of L2 writing ability.

Experimental studies need to systematically manipulate the type and amount of feedback provided in each treatment condition so that any differences between treatment groups can be attributed solely to the manipulated (independent) variable instead of extraneous, confounding variables. A finding that GenAI feedback is more effective than teacher feedback is difficult to interpret if GenAI and teachers provided different types and amounts of feedback. GenAI can provide different types of feedback, which previous research has found to have differential effects on learning gains (Kang and Han 2015; Li and Vuono 2019). Experimental research is needed to examine the extent to which the results can be replicated in GenAI contexts, and to investigate new variables or opportunities created by GenAI such as integrated versus detached feedback. Theoretically, integrated feedback, which is embedded in the text, is contextualized and provides an immediate comparison between errors and correct forms. Thus, it is more effective than detached feedback, which takes the form of a list of corrections at the end of the text. Another example topic with a

significant potential in GenAI contexts is the timing of corrective feedback (Li 2020; Li et al. 2025). Feedback timing can be operationalized in various ways, such as focusing on different aspects in different drafts during the cycle of a writing assignment or correcting errors while learners are engaged in the process of composing versus after the writing task is completed (Cheng and Zhang 2024; Shintani and Aubrey 2016).

One area that appears related to but actually separate from feedback is writing assessment; the former focuses on helping students learn and the latter on evaluating writing quality or performance. One recurring theme of the research is that GenAI ratings are more consistent with human ratings when assessment criteria are clearly defined and sufficiently illustrated and when enough training is provided to GenAI. Research is needed on ways to increase the validity of GenAI assessments, and principles of effective assessments should be applied in the examination of the construct validity of GenAI assessments, such as reliability, concurrent validity, divergent validity, convergent validity, and predictive validity. The findings of this stream of research may provide valuable implications for large-scale standardized assessments as well as curriculum-based or classroom assessments.

Finally, research on the comparison of the discourse and linguistic features of GenAI and human writing shows that GenAI writing is more formal, rigid, impersonal, academic, linguistically complex, and formulaic, whereas human writing is more personal, creative, engaging, and less predictable. These findings help us distinguish human and GenAI writing, contributing to more accurate detection of GenAI use in plagiarism inspection and enabling us to make informed decisions in our attempts to maintain our writing identity and our identity as a unique human being.

To conclude this article, I propose three principles that may guide the use, research, and theory construction of GenAI. The first and overarching principle is individualization. Individualization may be understood as the personalization of writing style including content, language, organization, and other aspects of writing. Personalization contrasts with commonality, patterning, and typicality – defining characteristics of GenAI output; therefore, personalization sets human writing apart from GenAI writing. As reported in previous sections, GenAI writing has been found to lack personal (authorial) voice, creativity, and logical reasoning – characteristics of human writing. Another way to apply the concept of individualization is to tailor GenAI output to fit user characteristics such as L2 writers' proficiency level, age, personal traits and dispositions (anxiety, motivation, working memory, etc.), and so on. Individualized assistance is particularly important for GenAI use because GenAI does not provide tailored assistance and it provides the same output to all users without considering their individual differences.

The second principle is domain-specificity. Here the scope of domain varies across contexts and may refer to an area of learning such as L2 learning, a subarea of learning such as L2 writing, and a particular task within the subarea. Some examples may help illustrate the importance of domain-specificity. For example, using GenAI to assess TOEFL writing involves training GenAI using the TOEFL rating rubric and scored TOEFL essays. Asking GenAI to compose a text necessitates the provision of task-specific information regarding the goal, genre, audience, assumed writer's identity, etc. in addition to the exact writing prompt.

The third principle is agency, which means that users should take a critical stance vetting the quality of GenAI and exercising autonomy instead of overreliance on GenAI. Users should take control over the whole process by actively improving GenAI literacy, learning or using strategies for effective prompt engineering, analyzing GenAI output to determine its acceptability instead of passively accepting it, and reflecting on the process for future improvement. Agency also means maintaining one's identity when taking advantage of GenAI output, such as determining whether a change suggested by GenAI fits one's own style or makes one modify their writing in a dramatic way to the extent that their identity is lost.

**Research ethics:** Not applicable.

**Informed consent:** Not applicable.

**Author contributions:** The author has accepted responsibility for the entire content of this manuscript and approved its submission.

**Use of Large Language Models, AI and Machine Learning Tools:** None declared.

**Conflict of interest:** The author states no conflict of interest.

**Research funding:** None declared.

**Data availability:** Not applicable.

## References

- An, H., and S. Li. 2024. "Task-Specific Writing Anxiety and Self-Efficacy are Separate from General L2 Writing Anxiety and Self-Efficacy and they have Differential Associations with the Effects of Written Corrective Feedback in Pre-Task and Within-Task Planning." *System*: 1–21. <https://doi.org/10.1016/j.system.2024.103480>.
- Boudouaia, A., S. Mouas, and B. Kouider. 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): 1509–37.
- Bucol, J., and N. Sangkawong. 2024. "Exploring ChatGPT as a Writing Assessment Tool." *Innovations in Education and Teaching International*: 1–17. First view. <https://doi.org/10.1080/14703297.2024.2363901>.
- Casal, E., and M. Kessler. 2023. "Can Linguists Distinguish between ChatGPT/AI and Human Writing?: A Study of Research Ethics and Academic Publishing." *Research Methods in Applied Linguistics* 2: 1–12.

- Cheng, X., and L. Zhang. 2024. "Investigating Synchronous and Asynchronous Written Corrective Feedback in a Computer-Assisted Environment: EFL Learners' Linguistic Performance and Perspectives." *Computer Assisted Language Learning*: 1–30. <https://doi.org/10.1080/09588221.2024.2315070>.
- Dang, A., and H. Wang. 2024. "Ethical Use of Generative AI for Writing Practices: Addressing Linguistically Diverse Students in U.S. Universities' AI Statements." *Journal of Second Language Writing* 66: 1–8.
- Ellis, R. 2009. "A Typology of Written Corrective Feedback Types." *ELT Journal* 63: 97–107.
- Escalante, J., A. Pack, and A. Barrett. 2023. "AI-Generated Feedback on Writing: Insights into Efficacy and ENL Student Preference." *International Journal of Educational Technology in Higher Education* 20: 1–20.
- Guo, K., and D. Wang. 2024. "To Resist it or to Embrace it? Examining ChatGPT's Potential to Support Teacher Feedback in EFL Writing." *Education and Information Technologies* 29: 8435–63.
- Han, J., and M. Li. 2024. "Exploring ChatGPT-Supported Teacher Feedback in the EFL Context." *System* 126: 1–11.
- Huang, L., and J. Deng. 2025. "'This Dissertation Intricately Explores...': ChatGPT's Shell Noun Use in Rephrasing Dissertation Abstracts." *System* 129: 1–16.
- Jiang, Y., J. Hao, M. Fauss, and C. Li. 2024. "Detecting ChatGPT-Generated Essays in a Large-Scale Writing Assessment: Is There a Bias against Non-Native English Speakers?" *Computers & Education* 217: 1–14.
- Jiang, F., and K. Hyland. 2024. "Does ChatGPT Argue like Students? Bundles in Argumentative Essays." *Applied Linguistics*: 1–17. First view. <https://doi.org/10.1093/applin/amae052>.
- Kang, E., and Z. Han. 2015. "The Efficacy of Written Corrective Feedback in Improving L2 Written Accuracy: A Meta-Analysis." *The Modern Language Journal* 99: 1–18.
- Kim, H., and M. Bowles. 2019. "How Deeply Do Second Language Learners Process Written Corrective Feedback? Insights Gained from Think-Alouds." *Tesol Quarterly* 53: 913–38.
- Kim, J., and S. Li. 2024. "The Effects of Task Repetition and Corrective Feedback on L2 Writing Development." *Language Learning Journal*: 1–16. <https://doi.org/10.1080/09571736.2024.2390555>.
- Koltovskaia, S., P. Rahmati, and H. Saeli. 2024. "Graduate Students' Use of ChatGPT for Academic Text Revision: Behavioral, Cognitive, and Affective Engagement." *Journal of Second Language Writing* 65: 1–15.
- Lee, I. 2004. "Error Correction in L2 Secondary Writing Classrooms: The Case of Hong Kong." *Journal of Second Language Writing* 285–312. <https://doi.org/10.1016/j.jslw.2004.08.001>.
- Lee, I. 2008. "Student Reactions to Teacher Feedback in Two Hong Kong Secondary Classrooms." *Journal of Second Language Writing* 17: 144–64.
- Li, S. 2017. "Teacher and Learner Beliefs About Corrective Feedback." In *Corrective Feedback in Second Language Teaching and Learning*, edited by H. Nassaji, and E. Kartchava, 143–57. Milton Park: Routledge.
- Li, S. 2020. "What is the Ideal Time to Provide Corrective Feedback? Replication of Li, Zhu & Ellis (2016) and Arroyo & Yilmaz (2018)." *Language Teaching* 53 (1): 96–108.
- Li, S., and M. Fu. 2018. "Strategic and Unpressured Within-Task Planning and Their Associations with Working Memory." *Language Teaching Research* 22: 232–53.
- Li, S., L. Ou, and I. Lee. 2025. "The Timing of Corrective Feedback in Second Language Learning." *Language Teaching*: 1–17. First view. <https://doi.org/10.1017/S0261444824000478>.
- Li, S., and S. Roshan. 2019. "The Associations between Working Memory and the Effects of Four Different Types of Written Corrective Feedback." *Journal of Second Language Writing* 45: 1–15.
- Li, S., and A. Vuono. 2019. "Twenty-five Years of Research on Oral and Written Corrective Feedback." *System* 84: 93–109.
- Liang, W., M. Yuksekgonul, Y. Mao, E. Wu, and J. Zou. 2023. "GPT Detectors are Biased against Non-Native English Writers." *Patterns* 4: 1–4.

- Lin, S., and P. Crosthwaite. 2024. "The Grass is not Always Greener: Teacher vs. GPT-Assisted Written Corrective Feedback." *System* 127: 1–19.
- Mao, Z., I. Lee, and S. Li. 2024. "Written Corrective Feedback in Second Language Writing: A Synthesis of Research in Naturalistic Classroom Contexts." *Language Teaching*: 1–29. <https://doi.org/10.1017/S0261444823000393>.
- Mizumoto, A., N. Shintani, M. Sasaki, and F. Teng. 2024. "Testing the Viability of ChatGPT as a Companion in L2 Writing Accuracy Assessment." *Research Methods in Applied Linguistics* 3: 1–15.
- Ng, D. T. K., J. K. L. Leung, S. K. W. Chu, and M. S. Qiao. 2021. "Conceptualizing AI Literacy: An Exploratory Review." *Computers and Education: Artificial Intelligence* 2: 1–17.
- Nguyen, L., P. Dao, and B. Nguyen. 2024. "Model Texts as a Feedback Instrument in Second Language Writing: A Systematic Review." *Language Teaching Research*: 1–24. <https://doi.org/10.1177/136216882412913>.
- Shin, D., and J. Lee. 2024. "Exploratory Study on the Potential of ChatGPT as a Rater of Second Language Writing." *Education and Information Technologies*: 1–23. First view. <https://doi.org/10.1007/s10639-024-12817-6>.
- Shintani, N., and S. Aubrey. 2016. "The Effectiveness of Synchronous and Asynchronous Written Corrective Feedback on Grammatical Accuracy in a Computer-Mediated Environment." *Modern Language Journal* 100 (1): 296–319.
- Steiss, J., T. Tate, S. Graham, J. Cruz, M. Hebert, J. Wang, et al. 2024. "Comparing the Quality of Human and ChatGPT Feedback of Students' Writing, Learning and Instruction." *Learning and Instruction* 91: 1–15.
- Tan, X., C. Wang, and W. Xu. 2025. "To Disclose or not to Disclose: Exploring the Risk of Being Transparent About GenAI Use in Second Language Writing." *Applied Linguistics*: 1–15. First view. <https://doi.org/10.1093/applin/amae092>.
- Truscott, J. 1996. "The Case against Grammar Correction in L2 Writing Classes." *Language Learning* 46: 327–69.
- Tu, S. 2025. "Exploring ChatGPT's Potential as an AI-Powered Writing Assistant: A Comparative Analysis of Second Language Learner Essays." *Language Teaching*: 1–3. First view. <https://doi.org/10.1017/S0261444824000259>.
- Voss, E., and Z. Waring. 2024. "When ChatGPT Can't Chat: The Quest for Naturalness." *Tesol Quarterly*: 1–12. First view. <https://doi.org/10.1002/tesq.3374>.
- Vuogan, A., and S. Li. 2023. "The Effectiveness of Peer Feedback in L2 Writing: A Meta-Analysis." *Tesol Quarterly* 57: 1115–38.
- Wang, C., and Z. Wang. 2025. "Investigating L2 Writers' Critical AI Literacy in AI-Assisted Writing: An APSE Model." *Journal of Second Language Writing* 67: 1–17. First view.
- Warschauer, M., W. Tseng, S. Yim, T. Webster, S. Jacob, Q. Du, et al. 2023. "The Affordances and Contradictions of AI-Generated Text for Writers of English as a Second or Foreign Language." *Journal of Second Language Writing* 62: 1–17.
- Yamashita, T. 2024. "An Application of Many-Facet Rasch Measurement to Evaluate Automated Essay Scoring: A Case of ChatGPT-4.0." *Research Methods in Applied Linguistics* 3: 1–14.
- Zhang, M., and P. Crosthwaite. 2025. "More Human Than Human? Differences in Lexis and Collocation within Academic Essays Produced by ChatGPT-3.5 and Human L2 Writers." *International Review of Applied Linguistics*: 1–28. First view. <https://doi.org/10.1515/iral-2024-0196>.
- Zou, S., K. Guo, J. Wang, and Y. Liu. 2025. "Investigating Students' Uptake of Teacher- and ChatGPT-Generated Feedback in EFL Writing: A Comparison Study." *Computer Assisted Language Learning*: 1–30. First view. <https://doi.org/10.1080/09588221.2024.2447279>.