INNOVATION IN LANGUAGE LEARNING International Conference

From Perceptions to Practice: Exploring EFL University Teachers' Prompt Literacy in Generative AI-Assisted Language Teaching

Hessah Saleh Aldayel

English Language Skills Department, Common First Year Deanship, King Saud University, Riyadh, Saudi Arabia

Abstract

Generative artificial intelligence (GenAI) is becoming part of everyday English language teaching, and its usefulness depends largely on teachers' ability to write effective prompts, known as prompt literacy [6]. EFL teachers generally view GenAI positively [11] and ask for training in writing effective prompts [2], yet research on prompting has relied largely on teachers' self-reported perceptions rather than the prompts they actually produce, although what teachers report does not always match what they do [3]. This convergent mixed-methods study therefore examined whether what teachers say about their prompting matches what they write. All of the approximately 200 teachers in the English department serving foundation-year students at a Saudi university were invited, and 20 completed a three-part online instrument: a questionnaire on GenAI use and self-perceived prompt literacy, two tasks in which teachers wrote prompts to obtain GenAI feedback on student paragraphs, and open-ended questions on the challenges of writing prompts. Prompts were scored by two independent raters with a researcher-developed checklist adapted from Li's input–task–output framework [9] (weighted kappa = .87), and open-ended responses were analysed thematically. Teachers rated their own prompt writing highly, but their prompts scored low (mean 6.6 of 18), and self-perception showed no positive relationship with prompt quality ($\rho = -.30$): teachers stated clear actions but rarely provided context, purpose, or output specifications. The findings offer an early performance-based account of in-service EFL teachers' prompt literacy, a tool that other researchers can adapt, and a clear target for prompt-focused teacher development.

Keywords: *prompt literacy; EFL teachers; professional development; teacher cognition; AI in language teaching and learning*

1. Introduction

Generative artificial intelligence (GenAI) systems, such as ChatGPT, are increasingly being integrated into English language teaching, where teachers use them to plan lessons, give feedback, and develop materials. In line with the digital transformation goals of Saudi Vision 2030, these tools are increasingly present in Saudi higher education, including foundation-year English programmes, where teachers work with large classes, an intensive curriculum, and heavy feedback loads, precisely the conditions under which GenAI support is most attractive. However, the usefulness of these tools is not automatic. Because GenAI systems respond to natural-language instructions, what a teacher obtains from them depends largely on the teacher's ability to write effective prompts, a skill known as prompt literacy [6]; indeed, the quality of prompts predicts the quality of the output that GenAI systems produce [7].

Although previous studies have examined EFL teachers' views of GenAI and prompting, including in Saudi Arabia, most have focused on teachers' perceptions rather than on the prompts they actually write. Teachers generally report positive attitudes toward GenAI [1], [11] and ask for training in writing effective prompts [2], and a recent review confirmed that most GenAI studies in language classrooms rely on self-reported data [8]. Yet what teachers say does not always match what they do [3], and what teachers believe they can do with technology often differs from what they can actually do [10], [5]. It therefore remains unclear whether teachers' self-perceived prompt literacy matches the quality of the prompts they write. The answer to this question can show what prompt-focused training should teach and which teachers need it most.

The present study addresses this question in the English department serving all foundation-year students at a university in Saudi Arabia. Using a convergent mixed-methods design, it compares what teachers say about their prompting (a questionnaire) with what they actually write (two prompt-writing tasks scored against a researcher-developed checklist adapted from Li's [9] input–task–output framework), and draws on open-ended answers to explain the results. In doing so, it offers an early

performance-based account of in-service EFL teachers' prompt literacy, an instrument that other researchers can adapt, and practical guidance for prompt-focused teacher development.

2. Literature Review

2.1 Prompt Literacy as an Emerging Teacher Skill

Prompt literacy is a term that appeared after the release of ChatGPT. Hwang et al. [6] introduced it in an exploratory study of language learners, describing it as a new literacy skill: the ability to write precise prompts for GenAI, interpret the output, and improve the prompts until they produce the intended result. Later studies developed the idea in different ways: Cain [4] shifted attention from the prompt itself to the thinking behind it, namely content knowledge, critical thinking, and iterative design; Walter [13] placed prompt engineering, together with AI literacy and critical thinking, among the main skills that schools and universities must now teach; and Tour and Zadorozhnyy [12] brought the idea back to language education, describing prompt literacy as the set of abilities needed to interact with GenAI and improve prompts, because the quality of a prompt directly shapes the relevance and accuracy of what the AI produces. Although these studies focused mainly on learners, teachers need the same skill whenever they use GenAI to plan lessons, create materials, or generate feedback on student writing. Across these studies, prompting is treated as a linguistic and cognitive activity (providing the input, defining the task, and specifying the output) rather than a list of tricks to memorise, and this view underpins the checklist used in the present study (see Section 2.4).

2.2 GenAI and Prompting in EFL Teaching: Research Built On Teachers' Perceptions

Research on GenAI in EFL teaching has grown quickly [8], including in Saudi Arabia, where the present study was conducted. Studies of Saudi EFL university teachers generally report positive views of GenAI for designing materials, planning lessons, and giving feedback, together with concerns about accuracy, ethical issues, and over-dependence on AI tools [1]. Bhatti [2] is the closest to the present research: surveying 77 EFL teachers across Saudi institutions and interviewing ten of them, that study found that teachers valued AI-generated content but that its usefulness depended on their ability to write effective prompts matched to their teaching goals, and they emphasised the need for training in this skill.

Despite these findings, all of these studies share one limitation: they rely on what teachers say rather than on what they actually do. Questionnaires and interviews record teachers' perceptions of their prompting, but they do not show the prompts teachers actually write. Research that has examined real prompts has focused almost entirely on students. Hwang et al. [6] analysed how language learners prompted GenAI, and Woo et al. [14] examined how secondary school EFL students used prompts when writing short stories with AI. Teachers' prompts have not received the same attention. Lee et al. [8], reviewing 49 GenAI studies in language classrooms, likewise found that most relied on self-reported data and called for research on what people actually do with GenAI. Since every use of GenAI begins with writing a prompt, the present study answers this call by examining the prompts teachers actually write.

2.3 From What Teachers Say to What They Do

A long-standing issue in language teacher education is the distinction between what teachers say about their teaching and what they really do in the classroom. Borg's [3] review of teacher cognition research shows that teachers' stated beliefs, knowledge, and self-perceptions do not necessarily translate into classroom practice, so understanding teaching requires looking at both what teachers say and what they do.

Research on teachers' digital competence (their skills in using technology) shows the same problem: studies usually ask teachers to rate their own skills, even though these self-ratings often fail to show what teachers can actually do. Maderick et al. [10] compared preservice teachers' self-ratings with objective measures across seven domains of digital competence and found that self-ratings alone were not an accurate measure of what these teachers could actually do, and González-Mujico [5] likewise recommends adding observed task performance to self-perceptions.

If beliefs and actual ability can differ for digital skills in general, the same gap is possible for prompt literacy, but it has not yet been examined. Knowing whether, and for which teachers, beliefs about prompting match the quality of the prompts they write has practical value: teachers who believe they are better at prompting than they actually are may see little need for training, even though they may need it

most; and if teachers' beliefs prove accurate, future researchers can simply use questionnaires, which are far easier than collecting and scoring real prompts. Either way, the answer shows who needs training and whether what teachers say about their prompting can be trusted.

2.4 The Scoring Framework

Scoring teachers' prompts requires a framework whose elements can be seen in a single written prompt, rather than in a longer conversation with a GenAI system. For this reason, the present study adapts the prompt dimensions identified by Li [9] in the context of GenAI and second language writing. Li's framework describes prompt quality through three dimensions: input specification (the information the teacher gives the AI, such as the student paragraph, the learners' level, and the course context), task specification (what the teacher asks the AI to do, for example to give feedback on organisation rather than to rewrite the paragraph), and output specification (what the teacher requires of the AI's answer, such as its format, tone, and length). Together, these three dimensions cover the information that decides whether a prompt is likely to produce useful feedback. In each task in the present study, the teacher wrote one prompt for one student paragraph, and all three dimensions could be seen and scored in that single prompt rather than in any later conversation between the teacher and the AI. Section 5.4 describes how they were turned into a researcher-developed checklist.

3. Objectives of the Study

The study aimed to describe foundation-year EFL teachers' reported use of GenAI tools and their self-perceived prompt literacy; to evaluate the quality of the prompts they write to obtain GenAI feedback on two student paragraphs; to examine how closely their self-perceived prompt literacy matches the quality of those prompts, identifying groups of teachers by how closely their beliefs match their prompt scores; and to identify the challenges they encounter when writing prompts for GenAI, in order to inform practical, prompt-focused teacher development.

4. Research Questions

The study addressed the following questions:
Q1. To what extent do foundation-year EFL teachers report using GenAI tools in their teaching, and how do they perceive their own prompt literacy?
Q2. What is the quality of the prompts teachers write to obtain GenAI feedback on two student paragraphs?
Q3. How closely does teachers' self-perceived prompt literacy match the quality of the prompts they write?
Q4. What challenges do teachers encounter when writing prompts for GenAI in their language teaching?

5. Methodology

5.1 Study Design

The study used a convergent mixed-methods design: quantitative and qualitative data were collected at the same time through one online instrument, and the results were brought together when they were interpreted. The instrument had three parts: a questionnaire on GenAI use and self-perceived prompt literacy, two short prompt-writing tasks, and open-ended questions about the challenges teachers face when writing prompts. Because each prompt-writing task asked the teacher to write the prompt they would use to get GenAI feedback on one student paragraph, the study collected the real prompts teachers wrote and could compare, within the same group of participants, what teachers say about their prompting, what they actually write, and how they explain their practice.

5.2 Participants and Setting

The study was conducted in the English department serving all foundation-year students at a university in Saudi Arabia, and all of its approximately 200 in-service EFL teachers were invited to participate. The foundation-year context was chosen because teachers follow a common syllabus, teach students with similar proficiency levels, and regularly give written feedback on student paragraphs: the prompt-writing tasks are therefore close to what these teachers already do, and because teachers work in such similar

conditions, differences in their prompts are more likely to reflect prompting ability than their teaching situation. Participation was voluntary, and all responses were collected anonymously.
Of the approximately 200 teachers invited, 20 completed the survey, a response rate of about 10%. Table 1 summarises the respondents.

Table 1. Demographic and Professional Characteristics of Respondents (n = 20)

Characteristic	n
Gender	
Female	18
Male	2
Teaching experience	
1–5 years	2
6–10 years	7
11–15 years	3
16–20 years	5
21 or more years	3
Highest qualification	
Doctorate	10
Master's	9
Bachelor's	1
Training on using GenAI tools	
Yes	10
No	10
Training on writing prompts	
Yes	6
No	14

Because participation was voluntary and anonymous, the sample is self-selected: teachers who already use GenAI tools, or who feel comfortable with them, may be over-represented. This is taken up as a limitation in Section 7.

5.3 Data Collection Instrument

Data were collected through an online instrument with three parts, written in English and delivered on a secure online survey platform.

Part 1: Questionnaire on GenAI use and self-perceived prompt literacy. Section A recorded background information (five items); Section B asked how often and for what purposes teachers used GenAI in their teaching (three items); and Section C asked how teachers usually write prompts (twelve five-point Likert items): three items for each of the three checklist dimensions (providing the input, defining the task, and specifying the output; see Section 5.4) and three items on overall confidence and habits of improving prompts, so that the questionnaire asked about the same things the checklist scored and the two could be compared fairly. The questionnaire was reviewed by three experienced EFL colleagues to check that the items covered the right content and piloted with a small group of teachers from a similar programme to check that the items were clear; internal consistency was estimated with Cronbach's alpha.

Part 2: Prompt-writing tasks. Teachers wrote, word for word, the prompts they would use to get GenAI feedback on two student paragraphs. Every teacher received the same two paragraphs, written by the researcher to resemble typical foundation-year student writing, with the kinds of language errors common at this level, and reviewed by experienced foundation-year teachers to confirm that they were realistic. The two paragraphs were similar in length and difficulty but on different topics, so that differences between prompts reflected prompting ability rather than differences in the paragraphs, and the two task scores could be compared directly. Teachers wrote each prompt exactly as they would type it into a GenAI tool, with no time limit; they were asked not to use any GenAI tool while completing the survey and were not shown the scoring checklist, so that their prompts showed their usual way of writing prompts.

Part 3: Open-ended questions. Open-ended questions asked teachers to describe the challenges they face when writing prompts, how they decide whether GenAI feedback is good enough to use with their students, and what support would help them most to improve their prompting.

5.4 Prompt Scoring and Inter-Rater Reliability

Each prompt was scored with a researcher-developed checklist adapted from Li's [9] input–task–output prompt framework. The adaptation keeps Li's three dimensions (input, task, and output specification) and turns each one into three checklist items written to fit the two prompt-writing tasks, giving nine items in total, every one of which can be seen in the written prompt: under input specification, whether the prompt includes the student paragraph, states the learners' level, and gives the course or task context; under task specification, whether it states a clear action, says what the feedback should focus on, and states the teaching purpose of the feedback; and under output specification, whether it specifies the format, the tone, and how long or detailed the feedback should be. Each item was scored on a three-point scale (0 = absent, 1 = present but vague, 2 = present and specific), with short descriptors and examples of what counts as vague and what counts as specific, giving a maximum of 18 points for each task; the two task scores were examined both separately and combined.

The nine items are labelled I1–I3 for the input dimension, T1–T3 for task specification, and O1–O3 for output specification; these are the labels used in the scoring sheets and in Appendix A. The checklist was first piloted on eight prompts, during which a codebook of scoring rules was written down and the descriptors for two items were refined: for O1, a bare request for feedback counts as no format stated, and for I2, any mention of the learner (for example "a student") scores 1, while an explicitly stated level (for example "B1") scores 2. The finalised checklist and codebook are included in Appendix A. Both raters then scored all 40 prompts independently. Three prompts were excluded from scoring: one teacher did not provide a prompt for Task 1, and one teacher declined both tasks, stating an objection to using GenAI for student feedback (this response is examined in Section 6.4). Across the 333 paired item scores for the remaining 37 prompts, exact agreement between the two raters was 84.1%, and the quadratic weighted kappa, which credits near-misses more than large disagreements, was .87, above the .70 threshold set in advance. The 53 disagreeing item scores were resolved through discussion: 42 were settled by the codebook rules, and for 11 the second rater's written reasoning was accepted. The resulting consensus scores were used in all analyses; as a robustness check, the analyses were repeated using the second rater's independent scores alone, and the pattern of findings reported in Section 6 was unchanged.

5.5 Data Analysis

The analysis combined quantitative and qualitative methods in three steps.
The first step summarised the data, reporting averages and percentages to show how teachers used GenAI, how they rated their own prompt literacy, and how their prompts scored (Q1 and Q2); before this, the data were checked for missing answers and unusual values.
The second step answered Q3 by comparing each teacher's self-perception score (the mean of items 9–17, on a 1–5 scale) with their prompt score (the mean score per completed task, out of 18), using Spearman's rank correlation because only 19 teachers had scored prompts. Teachers were then divided into three groups using a boundary set before the grouping was carried out: both scores were converted to z-scores so that the two measures shared one scale, and a teacher whose self-perception z-score exceeded their prompt z-score by more than 0.5 standard deviations was placed in the rated-themselves-higher group, a difference of more than 0.5 in the other direction placed a teacher in the rated-themselves-lower group, and all remaining teachers counted as a match. Subgroups of teachers (for example, by years of teaching or by prior GenAI training) were compared descriptively, without significance tests, because the groups were small.
The third step was qualitative and answered Q4. The open-ended answers were analysed using Thematic Content Analysis (TCA): the answers were read several times, their main ideas were labelled, and the labels were grouped into themes. The second rater also checked the labelling of a sample of answers, and any differences were discussed until agreement was reached.
Finally, the three sets of results (the questionnaire answers, the prompt scores, and the themes) were brought together and interpreted, for example by examining whether teachers who rated themselves high but whose prompts scored low also described few challenges with prompting.

5.6 Ethical Considerations

Ethical approval was obtained from the Standing Committee for Research Ethics at King Saud University (Ref. No. KSU-HE-26-1159; approved 23 August 2026) before data collection began. Teachers received an information sheet explaining the purpose of the study, gave their consent online, and could stop at any point before submitting their answers; because responses were anonymous, answers could not be withdrawn after they had been submitted. No identifying information was collected: each teacher's answers and prompts were analysed under an anonymous code.
All data were stored on a password-protected device to which only the researcher had access. Participation was voluntary, involved minimal risk, and had no effect on teachers' work or on how they were evaluated at the university. The two student paragraphs used in Part 2 were not real students' work: they were written by the researcher to resemble typical foundation-year writing, so no student data were collected or used in the study.

6. Findings and Analysis

6.1 GenAI Use and Self-Perceived Prompt Literacy

Most respondents were regular users of GenAI tools: 11 of the 20 said they used them daily or almost daily, 5 about once a week, 3 about once a month, and 1 never. ChatGPT was the most widely used tool (17 teachers), followed by Gemini (12), Copilot (5), and Claude (3); single teachers also named Poe and NotebookLM, and one had used no tool. The most common purposes were creating materials or activities (16), creating quizzes or tests (15), and planning lessons (13); notably, generating feedback on student writing, the activity used in the prompt-writing tasks, was selected by fewer than half of the teachers (9).
Self-perception was high across the questionnaire. The twelve item means ranged from 3.50 to 4.25 on the 1–5 agreement scale, and the three dimensions were rated almost identically: input 3.80, task 3.90, and output 3.82, with overall confidence at 3.95. The self-perception score (the mean of items 9–17) averaged 3.87 (SD = 0.75). Internal consistency was good: Cronbach's alpha was .86 for items 9–17 and .88 for all twelve items. In short, these teachers described themselves as competent prompt writers across all three dimensions of the framework.

6.2 Quality of Teachers' Prompts

The scored prompts told a different story. Across the 37 scored prompts, the mean consensus score was 6.55 out of 18 (SD = 2.62), roughly one third of the available points, with teacher means ranging from 1 to 11.5. Scores were similar for the two tasks (Task 1 mean 6.8; Task 2 mean 6.4), and eight teachers submitted the same prompt text for both tasks.
The item-level profile shows exactly where prompts were strong and where they were thin. Almost every teacher stated a clear action (mean 1.86 of 2 on item T1), and most named at least a broad feedback focus (T2 mean 1.30). Every other element was largely absent: specifying the format of the response averaged 0.78, including or referring to the student paragraph 0.76, stating the learners' level 0.68, tone 0.35, length 0.32, the teaching purpose of the feedback 0.30, and course or task context 0.27. In plain terms, teachers told the tool what to do, but rarely who the feedback was for, why it was wanted, or what shape the answer should take.

6.3 How Closely Perception Matches Performance

The central comparison of the study is between what teachers said and what they wrote. Across the 19 teachers with scored prompts, the Spearman correlation between the self-perception score and the consensus prompt score was $\rho = -.30$, which is close to zero and, if anything, negative. Teachers who rated themselves more highly did not write better prompts. The robustness check using the second rater's independent scores gave the same picture ($\rho = -.37$).
Using the pre-set boundary of half a standard deviation, 6 teachers rated themselves higher than their prompts scored, 8 matched, and 5 rated themselves lower (Table 2). The extreme cases make the pattern concrete. One teacher answered Strongly agree or Agree to every self-perception item, yet submitted the two-word prompt "grammar mistakes", scoring 1 of 18, the largest gap in the sample. At the other end, one teacher whose self-ratings were among the more modest wrote the highest-scoring prompts in the sample (11.5 of 18), including named error types and an explicit learning purpose. A third

teacher, among the most experienced in the sample and with training in both GenAI use and prompt writing and daily GenAI use, rated themselves Agree or above throughout, but did not attempt Task 1 and scored 4 of 18 on Task 2.

The six teachers with prompt-writing training scored somewhat higher on average than the thirteen without it (7.5 versus 6.1 of 18); with groups this small the comparison is descriptive only, but its direction is consistent with the training need suggested by the item profile above.

6.4 Challenges in Writing Prompts

The open answers about challenges fell into three themes. The first, and most common, was making prompts complete and specific enough: “Ensuring everything I need is within the prompt”; “you need to anticipate everything you need and because AI doesn’t think”; “I guess I do not detail it enough … I usually have to continue the conversation until I get what I want”. This theme matches the scored prompts directly: the elements teachers found hard to include (context, purpose, output details) are the same elements the checklist found missing.

The second theme was the tool misunderstanding the request or producing unreliable answers: “sometimes AI does not fully understand what you need exactly”; “it often gives me its own opinion”; “their use for student feedback tends to be overly positive”. Teachers in this theme located the problem in the tool rather than in the prompt, even though several of their own prompts left out the information the tool would have needed.

The third theme was seeing no challenges at all. Four teachers wrote “nothing”, “none”, or similar, and this group included the lowest-scoring teacher in the sample, whose three answers were “nothing”, “do not know”, and “nothing”, as well as the teacher who declined the prompt-writing tasks and answered “None. Writing prompts is easy … None, I have a brain.” Seeing no difficulty was not a sign of strong performance; in this sample it accompanied the weakest.

When asked how they judge whether GenAI feedback is good enough to use, most teachers pointed to their own expertise (“years of experience”, “I use my own judgement”, cross-checking the feedback against the paragraph) or to fit with their purpose and targets. Asked what support would help, the largest group asked for practical, hands-on training (“practical examples and hands-on practice rather than theory”, “prompt engineering training”, workshops and PD sessions), while three teachers suggested ready-made or standardised prompts that could be saved and reused, and the no-challenge group asked for nothing.

Table 2. How Closely Self-Perceived Prompt Literacy Matches Prompt Scores (n = 19)

Group	n (%)	Mean self-perception score	Mean prompt score
Self-perception higher than prompt score	6 (31.6%)	4.41	3.8
Self-perception matches prompt score	8 (42.1%)	3.89	6.9
Prompt score higher than self-perception	5 (26.3%)	3.20	9.2

7. Conclusion

This study set out to compare what EFL university teachers say about their prompt writing with the prompts they actually write. Most of the surveyed teachers used GenAI tools regularly and rated their own prompt writing highly and consistently, yet their written prompts scored low against the researcher-developed checklist, about one third of the available points, with a characteristic profile: clear actions, broad focus, and little else. Self-perception did not match performance: the correlation was close to zero, and nearly a third of the teachers rated themselves markedly higher than their prompts scored. The challenges teachers named mirrored the scored gaps for many, while a small group saw no challenges at all, a group that included the weakest performance in the sample.

The practical implication for professional development is specific. Training should not aim at ‘using AI’ in general: teachers already state clear actions. What their prompts lacked, almost uniformly, was context (who the learner is, what the course and task are), purpose (what the feedback is for), and

output specification (format, tone, length). These are teachable, concrete habits, and the teachers themselves asked for exactly this kind of practical, hands-on training. The findings also carry a warning for voluntary training models: the teachers most in need of support may be the least likely to seek it, because they see no difficulty. The refusal case adds a further point: departments introducing GenAI feedback should expect, and make room for, principled disagreement.

The study's limitations follow from its design and scale. Twenty teachers from one department completed the survey (about 10% of those invited), and the sample is self-selected, so the findings describe this group and cannot be generalised to the department or beyond. The survey was completed online without supervision, so it cannot be ruled out that some teachers wrote prompts with outside help, although the low scores make this unlikely to have inflated the results. Prompts were collected on a single occasion with two fixed tasks and scored with one researcher-developed checklist; other instruments might weight elements differently, and the subgroup comparisons are descriptive because the groups are small. Within these limits, the study offers an early performance-based account of teachers' prompt literacy, a checklist and codebook that other researchers can adapt, and a clear, data-based target for prompt-focused professional development. Future research should repeat the comparison with larger samples, in other departments and countries, and test whether the training suggested here actually closes the gap between perception and practice.

REFERENCES

[1] Alzubi, A. A. F., & Alelaiwi, A. S. (2025). Exploring EFL university teachers' perceptions of AI-generative tools in Saudi Arabia: A mixed-methods study. *Discover Computing*, *28*(1), Article 282. https://doi.org/10.1007/s10791-025-09793-x

[2] Bhatti, A. (2026). Strategic prompt engineering for enhancing AI-generated content in English language teaching empowering EFL contexts. *International Journal of Computer-Assisted Language Learning and Teaching*, *16*(1), 1–30. https://doi.org/10.4018/ijcallt.398504

[3] Borg, S. (2003). Teacher cognition in language teaching: A review of research on what language teachers think, know, believe, and do. *Language Teaching*, *36*(2), 81–109. https://doi.org/10.1017/s0261444803001903

[4] Cain, W. (2024). Prompting change: Exploring prompt engineering in large language model AI and its potential to transform education. *TechTrends*, *68*(1), 47–57. https://doi.org/10.1007/s11528-023-00896-0

[5] González-Mujico, F. de L. (2024). Measuring student and educator digital competence beyond self-assessment: Developing and validating two rubric-based frameworks. *Education and Information Technologies*, *29*(11), 13299–13324. https://doi.org/10.1007/s10639-023-12363-7

[6] Hwang, Y., Lee, J. H., & Shin, D. (2023). What is prompt literacy? An exploratory study of language learners' development of new literacy skill using generative AI. arXiv preprint arXiv:2311.05373.

[7] Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, *6*, 100225. https://doi.org/10.1016/j.caeai.2024.100225

[8] Lee, S., Choe, H., Zou, D., & Jeon, J. (2026). Generative AI (GenAI) in the language classroom: A systematic review. *Interactive Learning Environments*, *34*(1), 335–359. https://doi.org/10.1080/10494820.2025.2498537

[9] Li, S. (2025). Generative AI and second language writing. *Digital Studies in Language and Literature*, *2*(1), 122–152. https://doi.org/10.1515/dsll-2025-0007

[10] Maderick, J. A., Zhang, S., Hartley, K., & Marchand, G. (2016). Preservice teachers and self-assessing digital competence. *Journal of Educational Computing Research*, *54*(3), 326–351. https://doi.org/10.1177/0735633115620432

[11] Nazim, M., & Alzubi, A. A. F. (2025). Empowering EFL teachers' perceptions of generative AI-mediated self-professionalism. *PLoS ONE*, *20*(6), e0326735. https://doi.org/10.1371/journal.pone.0326735

[12] Tour, E., & Zadorozhnyy, A. (2025). Conceptualizing and operationalizing prompt literacy for English language learners. *Journal of Adolescent & Adult Literacy*, *69*(3), e70020. https://doi.org/10.1002/jaal.70020

[13] Walter, Y. (2024). Embracing the future of artificial intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education*, *21*(1), Article 15. https://doi.org/10.1186/s41239-024-00448-3

[14] Woo, D. J., Guo, K., & Susanto, H. (2025). Exploring EFL students' prompt engineering in human–AI story writing: An activity theory perspective. *Interactive Learning Environments*, *33*(1), 863–882. https://doi.org/10.1080/10494820.2024.2361381

Appendix A. Prompt Scoring Checklist and Codebook

Each prompt is scored on the nine items in the table below (0 = absent; 1 = present but vague or general; 2 = present and specific; maximum 18 per task). A score of 0 is given when the element does not appear in the prompt.

Item	What it scores	Score 1 (general)	Score 2 (specific)
I1	Includes or clearly refers to the student paragraph	Refers to it ("this paragraph")	Paragraph pasted into the prompt
I2	States the learners' level	General ("a student", "beginner")	Explicit level ("B1", "B2")
I3	Gives course or task context	General ("for my class")	Specific course or task information
T1	States a clear action	Vague ("help with this")	Clear ("give written feedback on this paragraph")
T2	Specifies the feedback focus	Broad category or rubric pointer ("grammar")	Named sub-aspects ("subject–verb agreement")
T3	States the teaching purpose	General ("to help the student")	Specific ("feedback the student can use to revise")
O1	Specifies the format of the response (bare feedback request = 0)	General form word ("give comments")	Explicit structure or order
O2	Specifies the tone or style	General ("constructive")	Specific ("teacher voice not AI voice")
O3	Specifies the length or scope	General ("concise"; "do not rewrite")	Quantified ("not more than 3 sentences")

Codebook Rules Applied by Both Raters:

Rule	Description
1	Score only what is written; a score of 2 requires concrete words present in the prompt.
2	The whole submitted prompt is scored as one unit.
3	General words alone score 1, never 2.
4	One phrase feeds several items only if it genuinely satisfies each; learner identity is scored under I2 and course or task under I3, never both from one phrase.
5	Items I1–T3 describe the student's text and task; items O1–O3 describe the AI's answer. A phrase about the writing never scores an O item.
6	Constraints (e.g., "do not rewrite") count toward the item they limit.
7	Unfilled template placeholders (e.g., "level (X)") score 1.
8	Identical prompt texts receive identical scores.
9	Non-attempts and refusals are recorded as such and excluded from prompt statistics.