Imaginex International Journal of Interdisciplinary Research and Reviews
Volume 1, Issue 3, November – December 2026
ISSN: 3140-0094 (Online)
www.iijirr.com

Prompt-Literacy Training for ESL Teachers: An Exploratory Pre–Post Study of ChatGPT Prompt and Output Quality

M.R. Malvika[1], Gangavarapu Suvarna Lakshmi[2], Ramamuni Reddy Chada[3], Ravikiran G[4]

[1]PhD Scholar, Department of English Language Teaching, The English and Foreign Languages University, Hyderabad.
[2]Professor, Department of English Language Teaching, The English and Foreign Languages University, Hyderabad.
[3]Assistant Professor, Department of ESL Studies, The English and Foreign Languages University, Hyderabad.
[4]Junior Lecturer, Department of English, Government Junior College, Kodad.

Abstract

Teachers increasingly use Generative Artificial Intelligence (GenAI) tools such as ChatGPT to prepare lesson plans, worksheets, and test items, but the value of what these tools produce depends heavily on how the request is written. This exploratory study examined whether a prompt-literacy intervention changed the quality of English as Second Language (ESL) teachers' ChatGPT prompts and of the material ChatGPT generated from them. Ten ESL teachers from one undergraduate college in Chennai, India, were recruited by convenience sampling. Each used ChatGPT to produce a lesson plan and a grammar worksheet before and after the intervention. Prompt quality was scored on the Ask SMILE platform, and the outputs were rated by the researcher on five criteria on a five-point rubric. Mean prompt scores rose from 2.76 (SD = 0.47) to 4.48 (SD = 0.51), and mean output scores rose from 3.02 (SD = 0.47) to 4.32 (SD = 0.49). Both gains were statistically significant (prompts: $t(9) = 17.24$; outputs: $t(9) = 24.19$; both $p < .001$), and every teacher improved on both measures. In interviews, all ten teachers preferred the outputs they obtained after the intervention. The sample was small, there was no comparison group, the same task types were repeated on the same day, and the outputs were rated by one researcher who knew the test phase. Within these limits, the results suggest that a hands-on intervention can move teachers towards more specific, context-rich prompts.

Keywords: ChatGPT, ESL Teachers, Generative Artificial Intelligence, Prompt Engineering, Prompt Literacy, Teacher Professional Development

1. Introduction

Generative AI tools such as ChatGPT have spread quickly through education. In language teaching, they are used mainly to generate content, give feedback, and support teachers' work (Li et al., 2024; Yang & Li, 2024). What these tools produce, however, depends heavily on the prompt the user writes. A request that states the teaching context, the learners, the goal, the format, and any constraints usually yields more targeted material than a general one (Korzynski et al., 2023), and better prompting has been linked to better output (Knoth et al., 2024).

Prompting is therefore an important part of teachers' AI literacy. A longer prompt does not, however, guarantee correct or pedagogically useful output. The result also depends on the model, the nature of the task, and the information available to the model, and the teacher still has to judge whether what comes back is accurate and usable (Sabzalieva & Valentini, 2023).

These issues matter particularly in ESL teaching. Material has to suit the learners' proficiency, language background, and educational experience, as well as the curriculum and the conditions of the class. A generic request for a lesson plan may leave out several of these factors, whereas a context-rich prompt can state them directly.

This study examined how a prompt-writing intervention related to changes in ESL teachers' prompts, in the quality of the material ChatGPT produced from those prompts, and in the teachers' views of that material. With ten teachers from a single college and no comparison group, the study is exploratory rather than confirmatory.

2. Research Questions

RQ1: How does the quality of ESL teachers' ChatGPT prompts change after a prompt-literacy intervention, and do the prompts become more specific to the teaching context?

RQ2: How does the quality of the lesson plans and grammar worksheets that ChatGPT generates from these prompts, as rated on a five-criterion rubric, change after the intervention, and how do the teachers view these outputs?

3. Rationale and Literature Review

Research on AI chatbots such as ChatGPT in education has concentrated on four areas relevant to this study: the role of AI in teaching and learning, the importance of prompt writing for productive interaction with chatbots, gaps in teachers' preparation for using AI, and the value of these tools for English as a Second Language (ESL) teaching.

3.1 Generative AI in Language Education

Studies of ChatGPT in language teaching and learning have examined its use for creating materials, supporting writing, giving feedback, and interacting with learners. Two recent systematic reviews map this work. Li et al. (2024) reviewed 36 studies from ChatGPT's first year and found that it served mainly to support self-directed learning, generate content, and assist teachers' workflows. They called for broader and longer-term studies, more attention to stakeholders' perspectives, and closer evaluation of feedback quality. Yang and Li (2024) reviewed 44 studies of ChatGPT for second language learning. ChatGPT was used mainly for content generation, feedback, and teaching support; controlling the context and customising the output were the most common prompt patterns; and most studies involved small samples of college EFL learners. The reviewers also called for more pre-test/post-test experimental designs.

For ESL teachers, AI-generated content can be a useful starting point for planning lessons, worksheets, and tasks. Even well-written output, however, can contain factual errors, pitch the language at the wrong level, drift away from the lesson goals, or use examples that do not suit the class. Teachers still need to edit and adapt the material, and they need specific digital competencies to use such tools ethically and effectively (Kohnke et al., 2023).

3.2 Prompt Literacy and Contextualised Prompting

In this study, prompt literacy means the ability to write and evaluate instructions to a generative AI tool for a particular teaching purpose. In practice, it involves defining the task, the context, the target audience, the constraints, and the expected output; reviewing the result; and revising the prompt when the output misses the instructional goal. Korzynski et al. (2023) treat prompt engineering as a new digital competence and offer practical guidelines for writing prompts, and guidance written for language teachers recommends stating the purpose and style of the output, adding constraints, and giving examples (Peachey, 2023). The skill matters for output quality: among students, higher-quality prompting predicted better output from a large language model (Knoth et al., 2024).

Sawalha et al. (2024) analysed undergraduates' conversations with ChatGPT while they solved problems in an engineering course. Answers were more accurate when students reformulated a question or split it into several questions than when they pasted it in unchanged, and about a quarter of the students consistently pasted questions without modification. These findings show the value of analysing prompts. However, they come from students solving technical problems and may not transfer to ESL teachers planning lessons and worksheets. The present study offers a small case study of how teachers write prompts to produce teaching material.

3.3 Teacher AI Literacy and Professional Development

Teachers' AI literacy goes beyond knowing how to operate a tool. It includes understanding the benefits and risks of AI, making pedagogical judgements, evaluating output, and considering ethical questions. The UNESCO AI competency framework for teachers (Miao & Cukurova, 2024) sets out 15 competencies across five dimensions: a human-centred mindset, ethics of AI, AI foundations and applications, AI pedagogy, and AI for professional learning. The last two are the most relevant to prompt-focused training, whose aim is not to make prompts longer but to help teachers state clearly what they need for instruction.

A single intervention can introduce ideas and provide guided practice, but it is unlikely to change classroom practice by itself. The intervention in this study is therefore treated as an initial professional development experience. The study did not measure transfer to the classroom, implementation, or learner achievement.

3.4 Research Gap and Contribution

Generative AI has opened new possibilities for teaching, learning, and assessment, and teachers are increasingly encouraged to use it. There is, however, a gap between promoting these tools and using them well in real classrooms. Many teacher workshops rely on presentations and demonstrations and offer little hands-on practice in applying the tools to everyday teaching tasks.

Even simple tasks, such as producing a lesson plan, a worksheet, a classroom activity, or an assessment item, depend on the teacher's ability to write clear, context-relevant prompts. Weak prompt literacy can therefore limit the educational value of AI in practice. Professional development needs to do more than introduce the tools: it

should build prompt literacy as a core part of AI-related teacher education (Kohnke et al., 2023; Miao & Cukurova, 2024).

Studies that analyse prompts have so far involved mainly students (Knoth et al., 2024; Sawalha et al., 2024), and most studies of ChatGPT in second language learning have focused on college learners (Yang & Li, 2024). This study responds by examining a hands-on prompt-writing intervention for ESL teachers, using the pre-test/post-test design that recent reviews have called for.

3.5 Relevance to the ESL Context

Prompt literacy is especially important in ESL teaching because teachers must take account of learners' different levels of proficiency, communicative needs, and sociocultural backgrounds. ChatGPT offers a range of affordances for language teaching and learning (Kohnke et al., 2023). For teachers, these include creating context-relevant materials, designing communicative tasks, supporting interaction, and giving feedback on learners' language use. How well this works depends not only on access to the tools but also on whether teachers can write prompts that fit the learning goals, the learners' proficiency, and the teaching context. An interaction-oriented speaking task for A2 learners, for example, needs different instructions from an advanced academic discussion task. Without enough prompt literacy, teachers may receive output that is linguistically unsuitable, poorly contextualised, or pedagogically unsound.

AI can also help teachers create speaking, listening, reading, and writing activities. Using it well in ESL classes, though, means evaluating the generated content carefully, adapting it to the learners, and treating the technology as a support for the teacher's judgement rather than a substitute for it. Prompt literacy therefore helps ESL teachers use AI to create suitable learning opportunities.

4. Methodology

4.1 Research Design, Setting, and Participants

The study used a single-group pre-test/post-test design. The participants were ten ESL teachers who teach undergraduate students, many of them from rural backgrounds, at Dharmamurthi Rao Bahadur Calavala Cunnan Chetty's Hindu College, Chennai, Tamil Nadu, India. They were recruited by convenience sampling, and Table 1 summarises their profile. Because the sample was small and came from one institution, the results should be treated as exploratory.

The researcher designed and delivered the intervention and collected all participant data, and had no prior relationship with the participants. Before and after the intervention, the teachers completed a questionnaire that combined categorical items, five-point rating scales, yes/no questions, and open-ended questions. It collected demographic information (Table 1); the teachers' prompts for the lesson plans and grammar worksheets, together with ChatGPT's responses; and their views on prompt writing and AI-generated material, including self-rated prompting ability, satisfaction with the responses, prompt-refinement strategies, and perceived changes in response quality. The rating-scale and yes/no items are not analysed in this paper.

Table 1. Profile of the Participants (*N* = 10)

Characteristic	Details
Gender	Male 4, female 6
Age range (years)	27–50
Teaching experience (years)	3–25
Highest qualification	PhD
Prior use of ChatGPT	None to very rare

4.2 Ethics

Participation was voluntary. The researcher's institution has no research ethics committee, so formal ethics approval was not obtained; the study was approved by the researcher's supervisor. The questionnaire stated that participation was voluntary and that completing it indicated consent. Participants' data were anonymised for analysis and reporting, and the teachers are identified only as P1 to P10.

4.3 Intervention

The intervention lasted two hours and drew on guidance in Skrabut's (2023) *80 Ways to Use ChatGPT in the Classroom*. It aimed to develop the teachers' prompt literacy so that they could obtain relevant, accurate, and pedagogically appropriate responses from ChatGPT. It focused on how the wording of a prompt shapes the

response and on how prompts can be refined for English language teaching (ELT). The practice tasks centred on interaction-oriented speaking tasks, learner feedback, and the evaluation of AI-generated teaching materials. Appendix B sets out the stages, timings, and materials.

The intervention was practical. It combined facilitator demonstrations, group discussion, hands-on work with ChatGPT, peer review, and individual reflection. Teachers compared ChatGPT responses and refined their first prompts in line with their teaching goals.

The session opened with an activity in which teachers compared the responses to a vague prompt and to a specific prompt for a speaking activity. This showed them why a prompt should state who the learners are, what the lesson aims to achieve, what the task requires, and what output is expected. The researcher then introduced prompt literacy and the elements of an effective prompt using a five-component framework based on practitioner guidance shared on LinkedIn in 2024: Context, Command, Criteria, Constraints, and Composition (5C). Examples showed how each component could be built into an ELT prompt.

Next, in small groups, teachers improved weak prompts for speaking tasks, learner feedback, and classroom interaction. They ran both the original and the improved prompts in ChatGPT and compared the outputs. The groups then wrote prompts for interaction-based speaking activities, specifying the learners' proficiency level, the interaction goal, any limitations, and the output format. They were asked to consider whether the materials would encourage interaction, including turn-taking, clarification, negotiation, and decision-making.

In the final stage, teachers exchanged prompts and assessed them with a checklist prepared by the researcher, which covered accuracy, relevance to the task, appropriateness, instructional value, organisation, and clarity. They then revised their prompts in light of the peer feedback. The session ended with an individual exit ticket: each teacher improved a weak prompt, identified the features of a good prompt, and named one limitation of AI answers that teachers need to verify. The checklist criteria closely matched the rubric later used to rate the outputs (Appendix A); Section 6 considers the effect of this overlap.

4.4 Pre-Test and Post-Test Tasks

In both phases, each teacher used ChatGPT to (a) develop a lesson plan and (b) develop a grammar worksheet. The task types were the same in both phases, but the questions set were different in each phase. The post-test was held on the same day, immediately after the intervention.

Using the same task types makes the two phases comparable, but it also introduces a testing effect. Teachers had already seen the material ChatGPT produced at the pre-test, so their post-test prompts could improve simply because they were writing a similar prompt a second time. A short interval between the phases reduces the chance that outside events or maturation explain the change, but it does not remove this testing effect and may even strengthen it. Because the questions differed between the phases, differences in their difficulty may also have affected the scores. The design therefore cannot separate the effect of the intervention from these influences. These limitations are accepted in an exploratory study whose purpose is to identify trends for later controlled research. In addition, the intervention practised speaking tasks and feedback rather than lesson plans and worksheets, so the post-test shows how far the teachers transferred what they had learned to related tasks.

4.5 AI System and Response-Generation Conditions

Data were collected with the free version of ChatGPT, identified at the time as GPT-4o mini. Teachers used their own devices, and each task was run in a new chat. Because the free version could switch between models according to usage limits, the exact model behind each response could not be independently verified, and the free interface does not allow generation settings, such as temperature, to be set or recorded. Teachers could revise their prompts freely during the intervention, but at the post-test each teacher submitted one prompt per task and could not change it afterwards. Teachers used the model's first output without further iteration, at both the pre-test and the post-test.

4.6 Instruments and Measures

(a) Prompt quality. Prompts were scored on the Ask SMILE platform (Ask SMILE, n.d.), which uses a generative AI model to rate user inquiries. Its rubric draws on Bloom's taxonomy (Anderson & Krathwohl, 2001) and rates the critical thinking and inquiry complexity of a question on a five-level scale, from Level 1 to Level 5. Scores were assigned automatically by the platform and were not interpreted or changed by the researcher. Automated scoring keeps the procedure consistent. However, because the rubric was designed for inquiry questions, it reflects the cognitive level of a request more than its contextual detail; Ask SMILE scores are therefore treated as an indirect measure of prompt quality.

(b) Output quality. Each ChatGPT output was rated on a rubric developed for this study (Appendix A). The rubric has five criteria: accuracy; relevance to the task; contextual and learner appropriateness; completeness and instructional usefulness; and organisation and clarity. Each criterion is scored from 1 (very poor) to 5 (excellent). Before use, the rubric was reviewed and approved by two professors from a central university. Each teacher's

output score is the mean of the rubric scores across both tasks, on the rubric's 1–5 scale. All outputs were rated by the researcher, who knew whether each output came from the pre-test or the post-test.

(c) Teachers' views. After the post-test, the researcher held brief interviews with all ten teachers about the outputs they had obtained. Each interview lasted about 5 to 10 minutes and centred on two questions: the teacher's experience of the intervention, and any improvement they saw in the quality of ChatGPT's responses. The interviews were recorded, and the recordings were analysed thematically.

4.7 Data Analysis

Means and standard deviations were calculated for prompt and output scores at each phase. Because the same teachers took part in both phases, paired-samples *t*-tests compared the pre-test and post-test scores, with the significance level set at .05. The normality of the difference scores was checked with the Shapiro–Wilk test. Because rubric scores are ordinal and the sample was small, Wilcoxon signed-rank tests were run as a robustness check. Effect sizes were calculated as Cohen's d_z (the mean difference divided by the standard deviation of the differences) and as d_{av} (the mean difference divided by the average standard deviation of the two phases), which is easier to compare with other studies. Spearman's rank correlations were used to explore the association between prompt scores and output scores.

5. Results

5.1 Changes in Prompt and Output Scores

Table 2 shows each teacher's prompt and output scores before and after the intervention. Every teacher scored higher at the post-test on both measures.

Table 2. Prompt and Output Scores of Each Teacher Before and After the Intervention

	Pre-test		Post-test	
Teacher	**Prompt**	**Output**	**Prompt**	**Output**
P1	3.2	3.0	4.8	4.4
P2	2.8	2.8	4.0	4.0
P3	3.2	3.6	5.0	4.8
P4	2.8	2.8	4.0	3.8
P5	2.2	2.6	4.0	3.8
P6	2.2	2.4	4.0	3.8
P7	2.0	2.6	4.0	4.0
P8	2.8	3.2	5.0	4.8
P9	3.2	3.8	5.0	5.0
P10	3.2	3.4	5.0	4.8

Note. Prompt scores were assigned by the Ask SMILE platform; output scores are the researcher's ratings on the rubric in Appendix A. Both are on a 1–5 scale.

As Table 3 shows, the mean prompt score rose from 2.76 to 4.48, a gain of 1.72 points, and the mean output score rose from 3.02 to 4.32, a gain of 1.30 points.

Table 3. Descriptive Statistics for Prompt and Output Scores (*N* = 10)

Measure	Pre-test *M* (*SD*)	Post-test *M* (*SD*)	Mean gain
Prompt score	2.76 (0.47)	4.48 (0.51)	1.72
Output score	3.02 (0.47)	4.32 (0.49)	1.30

Both gains were statistically significant (Table 4). The difference scores did not depart significantly from normality (Shapiro–Wilk: prompts, $W = 0.86$, $p = .085$; outputs, $W = 0.91$, $p = .258$), and the Wilcoxon signed-rank tests gave the same result for both measures. The effect sizes are very large. They largely reflect how little the gains varied between teachers: prompt gains ranged from 1.2 to 2.2 points, output gains ranged from 1.0 to 1.6

points, and the standard deviation of the output gains was only 0.17. The effect sizes should therefore be read with caution, particularly for the output scores, which came from a single rater who knew the test phase.

Table 4. Paired Comparisons of Pre-Test and Post-Test Scores

Measure	Mean difference [95% CI]	$t(9)$	p	d_z	d_{av}	Wilcoxon p
Prompt score	1.72 [1.49, 1.95]	17.24	< .001	5.45	3.51	.002
Output score	1.30 [1.18, 1.42]	24.19	< .001	7.65	2.71	.002

Note. CI = confidence interval; d_z = mean difference divided by the standard deviation of the differences; d_{av} = mean difference divided by the average of the pre-test and post-test standard deviations. Wilcoxon p values are exact and two-tailed; all ten differences were positive on both measures.

5.2 Association Between Prompt and Output Scores

Teachers whose prompts scored higher also tended to receive higher-rated outputs, both at the pre-test (Spearman's $\rho = .89$, $p < .001$) and at the post-test ($\rho = .93$, $p < .001$). However, the size of a teacher's prompt gain was not significantly related to the size of the output gain ($\rho = .49$, $p = .155$). These correlations are exploratory. With ten teachers they are imprecise, and the output ratings may have been influenced by the rater's knowledge of the test phase.

5.3 Changes in the Content of Prompts

After the intervention, the prompts contained more information about the learners, the length of the class, and the teaching setting. Table 5 illustrates this change with one participant's pair of prompts for the lesson-plan task. The pre-test prompt read: "Give me a lesson plan for the poem 'The Solitary Reaper.'" It named the lesson content but gave no time limit, target audience, language level, or cultural background. The post-test prompt read: "Please provide a 45-minute lesson plan for first-year undergraduates, whose first language is Tamil. The class has 60 Indian ESL learners. The learners range from beginner to intermediate." It specified each of these elements.

Table 5. Context Elements in One Participant's Pre-Test and Post-Test Lesson-Plan Prompts

Element	Pre-test prompt	Post-test prompt
Lesson duration	Not stated	45 minutes
Learners	Not stated	First-year undergraduates
Class size	Not stated	60 learners
Language background	Not stated	Indian ESL learners whose first language is Tamil
Proficiency level	Not stated	Beginner to intermediate

5.4 Teachers' Views of the Outputs

In the interviews, all ten teachers said they preferred the outputs they had obtained after the intervention to those from before it, and they agreed that the later responses were more relevant because their prompts were more detailed. Three main themes emerged: prompt specificity, learner appropriateness, and alignment with instructional goals. Participants noted that refined prompts produced clearer, more relevant, and more goal-oriented materials suited to the learners' proficiency levels. One participant observed that "the lesson plan is specific, concise, and tailored well to the specific learners," while another said that "the result was goal-oriented and very apt to what I wanted." Overall, participants saw prompt refinement as improving the relevance and usefulness of ChatGPT-generated teaching materials.

6. Discussion

This exploratory study found that, after a two-hour intervention, ESL teachers wrote prompts that scored higher on Ask SMILE and that stated more about the learners, the lesson length, and the setting. Every teacher improved. This is consistent with work that treats prompt writing as a digital competence that can be developed (Korzynski et al., 2023) and with the finding that controlling the context and customising the output are the prompt patterns most often reported in studies of ChatGPT for language learning (Yang & Li, 2024). Because Ask SMILE rates the cognitive level of a request rather than its contextual detail, however, the prompt scores capture the gain in context-specificity only indirectly.

Output scores also rose, by 1.30 points on average. This fits evidence from studies with students, in which more developed prompting produced better or more accurate output (Knoth et al., 2024; Sawalha et al., 2024). Teachers whose prompts scored higher also received higher-rated outputs at both phases. The size of each teacher's prompt gain, however, did not predict the size of the output gain, so this study cannot show that better prompts caused better outputs.

The output gain (1.30 points) was smaller than the prompt gain (1.72 points). One likely reason is that a more specific prompt can improve how well material fits the learners and the task, but it cannot by itself correct factual or linguistic errors in what the model produces. Output quality also depends on the model and the task, and teachers kept the first output without revising their prompts, so they had no chance to correct weak responses.

Several other explanations must also be considered. First, the same task types were repeated on the same day, and teachers had seen their pre-test outputs, so part of the gain may reflect a second attempt rather than the intervention; the questions also differed between the phases, and differences in their difficulty may have affected the scores. Second, the researcher who led the intervention also rated the outputs and knew which phase each came from, which may have raised the post-test ratings. Third, the peer-review checklist used in the intervention matched the output rubric, so teachers learned the criteria on which their outputs would be judged. This alignment is reasonable in a training study, but it may have helped teachers ask for exactly the features the rubric rewards. Fourth, the teachers' stated preference for the post-intervention outputs may partly reflect a wish to please the trainer. Fifth, Table 1 shows that the teachers had rarely or never used ChatGPT before the study, so part of the gain may reflect first-time familiarity with the tool rather than the prompt training itself. The very large effect sizes should be read in this light.

Even with these limits, the results suggest that a short, hands-on intervention can move teachers from generic requests towards prompts that state the learners, the lesson length, and the context. This supports including prompt literacy in AI-related teacher development, alongside the AI pedagogy and AI for professional learning dimensions of the UNESCO framework (Miao & Cukurova, 2024). Better prompts do not replace the teacher's judgement, however. Teachers still need to check AI-generated material for accuracy and suitability before using it in class (Kohnke et al., 2023; Sabzalieva & Valentini, 2023).

7. Conclusion

This exploratory study examined ten ESL teachers' ChatGPT prompts, and the material generated from them, before and after a two-hour prompt-literacy intervention. Mean prompt scores rose from 2.76 to 4.48 and mean output scores from 3.02 to 4.32, and every teacher improved on both measures. After the intervention, the prompts also said more about the learners, the lesson length, and the setting. These results are encouraging, but they do not yet show that the intervention caused better material. The sample was small, there was no comparison group, the same task types were repeated on the same day, and the outputs were rated by one researcher who knew the test phase. Further research should use independent raters who are blind to the test phase, a comparison group, a delayed post-test, and a direct measure of how specific prompts are to the teaching context.

8. Limitations

- The sample consisted of ten teachers from one college, recruited by convenience sampling, so the findings cannot be generalised.
- There was no comparison group, so the effect of the intervention cannot be separated from practice, maturation, or expectation effects.
- The same task types were used at the pre-test and the post-test on the same day, and teachers had seen their pre-test outputs, which may have produced a testing effect. The questions also differed between the phases, so differences in their difficulty may have affected the scores.
- All outputs were rated by a single researcher who delivered the intervention and knew the test phase, so inter-rater reliability could not be calculated.
- Ask SMILE was designed to rate inquiry questions, and its validity as a measure of prompt quality has not been established.
- The researcher who led the intervention also conducted the interviews, so the teachers' views may be affected by social desirability.
- The model behind each response could not be independently verified, and the free interface does not allow generation settings to be controlled.
- There was no delayed post-test, so it is not known whether the change lasted.

9. Directions for Further Research

- Compare trained teachers with a comparison group who repeat the same tasks without training.

- Have the outputs rated by at least two independent raters who are blind to the test phase, and report inter-rater reliability.
- Measure the context-specificity of prompts directly, for example by coding prompts for the five 5C components.
- Add a delayed post-test to see whether improvements in prompt writing last.
- Analyse teachers' views of AI-generated material in more depth through systematic qualitative analysis.
- Test whether the gains extend to other teaching tasks, such as learner feedback and collaborative activities, and to other subjects.
- Study larger samples in other institutions, and follow teachers into the classroom to see how AI-generated material is used and whether it affects learner outcomes.

Declarations

Ethics approval and consent. Formal ethics approval was not obtained because the researcher's institution has no research ethics committee; the study was approved by the researcher's supervisor. Participation was voluntary, and consent was given through a statement on the questionnaire; completing it indicated consent.

Conflict of interest. The authors declare no conflict of interest.

Funding. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability. The anonymised scores are reported in Table 2. The prompts, ChatGPT outputs, and rubric score sheets are available from the corresponding author on reasonable request.

Use of generative AI. ChatGPT was the object of study and was used to generate the teaching materials analysed. During editorial review, the language of the manuscript was edited with the assistance of an AI tool (Claude, Anthropic); the authors reviewed and accepted the edited text.

References

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Ask SMILE. (n.d.). [Online question-evaluation platform]. https://ask.smile.seedsofempowerment.org/

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*, Article 100225. https://doi.org/10.1016/j.caeai.2024.100225

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

Korzynski, P., Mazurek, G., Krzypkowska, P., & Kurasinski, A. (2023). Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. *Entrepreneurial Business and Economics Review, 11*(3), 25–37. https://doi.org/10.15678/EBER.2023.110302

Li, B., Lowell, V. L., Wang, C., & Li, X. (2024). A systematic review of the first year of publications on ChatGPT and language education: Examining research on ChatGPT's use in language learning and teaching. *Computers and Education: Artificial Intelligence, 7*, Article 100266. https://doi.org/10.1016/j.caeai.2024.100266

Miao, F., & Cukurova, M. (2024). *AI competency framework for teachers*. UNESCO. https://doi.org/10.54675/ZJTE2084

Peachey, N. (2023). *ChatGPT in the language classroom*. Peachey Publications.

Sabzalieva, E., & Valentini, A. (2023). *ChatGPT and artificial intelligence in higher education: Quick start guide*. UNESCO International Institute for Higher Education in Latin America and the Caribbean. https://www.iesalc.unesco.org/wp-content/uploads/2023/04/ChatGPT-and-Artificial-Intelligence-in-higher-education-Quick-Start-guide_EN_FINAL.pdf

Sawalha, G., Taj, I., & Shoufan, A. (2024). Analyzing student prompts and their effect on ChatGPT's performance. *Cogent Education, 11*(1), Article 2397200. https://doi.org/10.1080/2331186X.2024.2397200

Skrabut, S. (2023). *80 ways to use ChatGPT in the classroom: Using AI to enhance teaching and learning*.

Yang, L., & Li, R. (2024). ChatGPT for L2 learning: Current status and implications. *System, 124*, Article 103351. https://doi.org/10.1016/j.system.2024.103351

Appendix A. Rubric for Evaluating the Quality of ChatGPT-Generated Materials

Evaluation criterion	1 – Very poor	2 – Poor	3 – Satisfactory	4 – Good	5 – Excellent
1. Accuracy	Contains major factual, linguistic, or pedagogical errors that undermine its usefulness.	Contains several errors or misleading explanations that require substantial correction.	Generally accurate but contains some errors, imprecisions, or questionable explanations.	Accurate overall, with only minor errors or imprecisions.	Consistently accurate in its content, explanations, examples, and pedagogical information.
2. Relevance to the task	Fails to address the requested task or its main objectives.	Addresses the task only partially and includes substantial irrelevant content.	Addresses the main task but includes omissions or some irrelevant material.	Addresses the task clearly, with most content directly supporting the stated objectives.	Directly and comprehensively fulfils the task, with all major elements closely aligned with the stated objectives.
3. Contextual and learner appropriateness	Ignores the specified learner profile, proficiency level, or teaching context.	Makes limited reference to the context, with substantial mismatches in level or suitability.	Partially accommodates the context but includes some unsuitable content or activities.	Appropriately reflects the learner profile, proficiency level, and teaching context, with minor mismatches.	Consistently and explicitly adapts the material to the specified learners, proficiency levels, and classroom context.
4. Completeness and instructional usefulness	Omits most essential components and cannot be used without being substantially recreated.	Includes few essential components and requires extensive additions or revision.	Includes the main components but requires several additions or modifications.	Includes nearly all necessary components and requires only minor modifications before use.	Includes all essential components needed for the intended task and is ready for classroom use with minimal or no modification.
5. Organisation and clarity	Disorganised, confusing, or difficult to interpret.	Poorly organised, with unclear instructions or substantial problems in sequencing.	Understandable overall but has some organisational weaknesses or ambiguous instructions.	Clearly organised, logically sequenced, and easy to follow.	Exceptionally clear, coherent, logically sequenced, and consistently easy for teachers and learners to follow.

Appendix B. Intervention Process

Stage	Activity	Duration	Example or material used	Skill targeted
1	Comparing vague and specific prompts	10 minutes	"Give me a speaking activity for students" compared with a detailed prompt specifying learner level, interactional objectives, duration, and output format	Recognising prompt specificity and its influence on AI responses
2	Introduction to prompt literacy	15 minutes	Examples of prompts illustrating context, instructional purpose, constraints, and output requirements	Understanding the principles of effective prompt formulation
3	Demonstration of the 5C prompt framework	20 minutes	Context, Command, Criteria, Constraints, and Composition; a sample prompt for an interaction-oriented speaking activity	Structuring prompts around pedagogical objectives and task requirements
4	Guided prompt-improvement exercise	20 minutes	Weak prompts for speaking activities, learner feedback, and interactive classroom tasks; ChatGPT-generated responses	Prompt revision, comparative evaluation, and iterative prompting
5	Break and reflection	10 minutes	Informal reflection on the importance of context and specificity in prompts	Consolidating learning
6	Application to ELT tasks and feedback	25 minutes	Prompts for a budget-friendly college-trip planning task and feedback on an anonymised learner interaction	Designing interaction-oriented tasks and generating pedagogically relevant feedback
7	Peer review and prompt refinement	15 minutes	Structured peer-review checklist and revised prompts	Critical evaluation, collaborative learning, and prompt refinement
8	Exit-ticket assessment	5 minutes	Individual task to revise "Make an interactive English lesson" and identify effective prompt features and AI limitations	Demonstrating individual understanding and application
Total		**120 minutes**		