Learning to Use AI for Learning: Teaching Responsible Use of AI Chatbot to K-12 Students Through an AI Literacy Module

Ruiwei Xiao[1], Xinying Hou[2], Ying-Jui Tseng[1], Hsuan Nieu[3], Guanze Liao[3], John Stamper[1], Kenneth R. Koedinger[1]
[1]Carnegie Mellon University
[2]University of Michigan
[3]National Tsing Hua University
ruiweix@cs.cmu.edu

Abstract

As Artificial Intelligence (AI) becomes increasingly integrated into daily life, there is a growing need to equip the next generation with the ability to apply, interact with, evaluate, and collaborate with AI systems responsibly. Prior research highlights the urgent demand from K-12 educators to teach students the ethical and effective use of AI for learning. To address this need, we designed a Large-Language Model (LLM)-based module to teach prompting literacy. This includes scenario-based deliberate practice activities with direct interaction with intelligent LLM agents, aiming to foster secondary school students' responsible engagement with AI chatbots. We conducted two iterations of classroom deployment in 11 authentic secondary education classrooms, and evaluated 1) AI-based auto-grader's capability; 2) students' prompting performance and confidence changes towards using AI for learning; and 3) the quality of learning and assessment materials. Results indicated that the AI-based auto-grader could grade student-written prompts with satisfactory quality. In addition, the instructional materials supported students in improving their prompting skills through practice and led to positive shifts in their perceptions of using AI for learning. Furthermore, data from Study 1 informed assessment revisions in Study 2. Analyses of item difficulty and discrimination in Study 2 showed that True/False and open-ended questions could measure prompting literacy more effectively than multiple-choice questions for our target learners. These promising outcomes highlight the potential for broader deployment and highlight the need for broader studies to assess learning effectiveness and assessment design.

1 Introduction

The increasing presence of AI in everyday life has amplified the need for effective techniques for human–AI interaction, including within educational settings (Van Brummelen, Shen, and Patton 2019). As AI products have increased adaptability and ease of use, students can access them easily in different contexts (Hou et al. 2025), including at home and school (Yim and Su 2024). Some AI commercial products, like ChatGPT, have been introduced to K-12 educational settings, aiming to personalize student learning experience (Zhang and Tur 2024; Nayak et al. 2023). Apart from its potential values, there are growing concerns about utilizing generative AI chatbot technologies including increasing cheating (Lee et al. 2024; Li et al. 2025; Xiao et al. 2025b) and the low quality of information retrieval in learning settings (Wang et al. 2023; Kazemitabaar et al. 2023).

While AI chatbots hold promise for K-12 education, few instructional interventions, to our knowledge, focus on teaching prompting literacy to K-12 students. Prompts are natural language instruction inputs that are used to communicate with AI chatbot technologies (Lo 2023). Prompting literacy includes helping students understand ***what*** AI can do to support learning, ***when*** to use AI effectively, and ***how*** to craft prompts for different types of assistance. While the potential of LLMs is promising, their current effectiveness remains heavily dependent on the quality of the prompts they receive (Ekin 2023). By designing and engineering them carefully, prompts can play a key role in shaping the AI's responses and guiding the output of LLMs to provide desired results (Lo 2023; White et al. 2023). As prompts become an important end-user communication channel with AI chatbots, scholars in various educational domains argue that the ability to create effective prompts that generate desired information is now an essential skill for students (Denny, Kumar, and Giacaman 2023; Korzynski et al. 2023; Meskó 2023; Woo, Guo, and Susanto 2024). Therefore, equipping students with prompting literacy is crucial to help students effectively communicate with AI chatbots during learning.

This work presents a web-based interactive instructional system to improve secondary-education (middle- and high-school) students' prompting literacy. It applies active learning (Bonwell and Eison 1991) and experiential learning (Ng et al. 2023) methods to engage students with prompting practice in three hypothetical learning scenarios. After a student writes a prompt, an AI auto-grader evaluates key dimensions and delivers immediate, detailed feedback. We hypothesized that practicing prompt writing in this way could help students learn to create effective prompts, supporting learning and confidence in AI use. We first outlined the module design. Then, we reported our classroom evaluation results. Specific research questions were listed under each evaluation study. Finally, we proposed ways to improve prompting literacy instruction.

2 Related Work

2.1 Prompting literacy in K-12 classrooms

It is believed that the ability to understand and use AI (or the lack thereof) will fuel the next digital divide in education (Trucano 2023). The emergence of generative AI makes such technologies more accessible to the general public and underscores that it is more urgent than ever to regulate the responsible and effective use of such technologies (Williamson, Molnar, and Boninger 2024). While K-12 teachers expressed their timely needs for AI competency materials for their students, especially prompting instructions (Lozano and Blanco Fontao 2023), the majority of current K-12 AI Literacy implementations emphasize understanding AI technologies than responsible use of AI. Additionally, existing AI competency frameworks primarily target adults in the workforce (Meskó 2023; Tseng et al. 2024; Zamfirescu-Pereira et al. 2023) and higher-education (Denny, Kumar, and Giacaman 2023; Giray 2023). As a focused area within AI literacy, prompting literacy has gained increasing attention since 2023 (Hwang, Lee, and Shin 2023; Lo 2023; Maloy and Gattupalli 2024; Dennison et al. 2024; Xiao et al. 2025a). However, the lack of emphasis on K-12 learners persists, as few studies specifically target secondary education students, despite growing concerns about student AI use in K–12 education (Li et al. 2025; Uanachain and Aouad 2025).

2.2 Learn-by-Doing with Elaborated Immediate Feedback

This study's instructional design is grounded in two foundational learning sciences principles: learning-by-doing (Schank, Berman, and Macpherson 2013) and elaborated immediate feedback (Wang et al. 2019). Learning-by-doing emphasizes active engagement through direct interaction with tasks. Prior research has shown that this approach is more effective than passive methods (e.g., reading, video-watching) in online learning environments (Koedinger et al. 2015). While learning-by-doing has been applied in prompting literacy education (e.g., Promptly (Denny et al. 2023)), such efforts often lack elaborated immediate feedback, which may hinder effective learning or even lead to dropout (Vasilyeva, De Bra, and Pechenizkiy 2008).

Recent advances in large language models (LLMs) offer new opportunities to address this challenge. LLMs have demonstrated strong performance in rubric-based grading across domains (Henkel et al. 2024; Nguyen et al. 2023) and are capable of generating personalized feedback for open-ended responses (Xiao, Hou, and Stamper 2024; Nguyen et al. 2023). For example, GPT-4 achieves up to 85% grading accuracy on short-answer questions using zero-shot prompting strategies (Henkel et al. 2024). Building on these capabilities, our work integrates an AI auto-grading pipeline with prompt-based practice to close the gap in prompting literacy instruction—enabling learning-by-doing accompanied by elaborated, immediate feedback to support more efficient and effective learning.

3 Prompting Literacy Module Overview

3.1 Learning Objectives

Based on the real needs of teachers (Lee et al. 2024) for instructions to teach students how to use AI for learning and the AI4K12 guideline (Touretzky et al. 2019), three learning objectives were derived: AI Capacity for Supporting Learning, Contexts to Use AI for Learning, and Effective Prompt Formation (Figure 2).

3.2 Learning Materials

We have designed and developed a web-based platform to support this prompting literacy practice. Students will first review the module's learning objectives, then complete two conceptual practice questions with feedback and hints to explore how AI can support their learning. After these, students will get into the main practice activities.

The main practice is conducted using three hypothetical learning scenarios. Each practice has four steps (Figure 1): 1) Introduce the detailed scenario to students, including the domain, the context, the student's current knowledge level, the learning goal the student is trying to achieve, and the resources the student has. 2) The student creates a prompt and clicks "generate" to receive answers from an AI chatbot. This mimics the authentic interaction with a general-purpose AI chatbot. 3) The student receives an answer from the AI chatbot based on the written prompt. 4) They can then click "Check my prompt" to receive an auto-graded result and a detailed explanation. This feedback is provided based on the preset evaluation dimensions for each scenario (Table 1). After receiving feedback, students can go back to craft a new prompt for the same practice or move on to the next practice. The detailed practice pipeline is shown in Figure 1.

The three scenarios were designed under three secondary school subjects: biology, geography, and math. Each domain was paired with a unique instructional activity, such as *"broaden knowledge beyond course requirements"*, *"prepare for a quiz tomorrow"*, and *"struggle with an assignment due tomorrow"*. Here is an example scenario:

> **Scenario 1: Use AI to extend your learning on biology**
> *Yesterday, you learned 'what is a cell' in the biology class. You roughly understood what cells are and the main components in a cell. Now, as you are interested in biology, you want to broaden your understanding of cells. This isn't about course requirements, just a topic (cell) you're interested in learning more about. The only resource you have is your biology textbook. Write a prompt to ask the AI Chatbot to help you extend your knowledge about cells!*

3.3 Automatic Grading Dimensions

In our study, the requirements for a high-quality prompt vary depending on the instructional activity the student is pursuing (Table 1). For example, a good prompt in the *homework struggle* situation should not explicitly ask for a direct answer. However, this criterion does not apply to other scenarios like *course content extension*. Therefore, when designing the criteria for a high-quality prompt, each practice for each scenario has specific dimensions to meet. Also, each dimension has its own in-context descriptions rooted in the

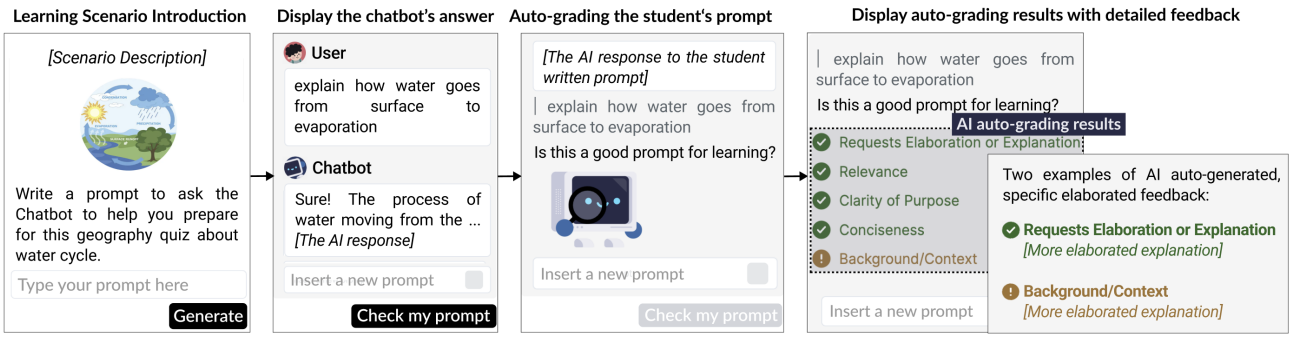

Figure 1: Students' practice pipeline in this interactive module

general definitions. In this study, we created our own grading dimensions based on the CLEAR framework for prompt creation (Lo 2023) and other existing prompting recommendations (Meskó 2023). Based on this grading rubric, for each student-written prompt, AI automatic grading needed to provide both a True/False categorical pass situation and a detailed explanation for this decision for each included dimension in that practice task. In addition, it was required to provide an detailed explanation for each dimension.

4 Study 1

As a novel activity type, we are interested in understanding students' perceptions of this activity, including what they learned from this activity, the interesting parts and the challenges encountered. In addition, as we integrated AI-based automatic grading into this procedure, we are interested in how it performed when grading students' written prompts using the preset evaluation dimensions. The first classroom evaluation study was conducted in June 2024 in 6 secondary-level classrooms in East Asia. We deployed this as an in-class practice with the local IRB (Institutional Review Board) approval. This resulted in valid data from 111 students. We explored the three RQs below:

- **RQ1:** What is the accuracy of AI-based automatic grading for students' written prompts?
- **RQ2:** How does this activity influence students' prompting ability and confidence in using AI for future learning?
- **RQ3:** How do students perceive the learning experience, including its benefits and challenges?

4.1 Study Procedure

The practice was conducted at Information Technology classes with a teacher to help facilitate the whole process. After a researcher explained the practice purpose to students, they first conducted a pre-practice survey about their prior experience with AI. Then the students completed six multiple-choice questions (pre-test) on prompting literacy and suitable/unsuitable learning scenarios with AI. After completing the pre-practice survey and tests, the student moved to practice their prompting literacy in the system. Teachers were instructed to inform students that the auto-grader may make mistakes.

After the practice, the student re-answered a 5-point Likert-scale survey question about *"I know how to use AI to help me learn"* and the six multiple-choice questions again as the post-test. At the end, the student answered 6 open-ended survey questions to reflect on the learning experiences. The LLM used was OpenAI GPT-4o.

4.2 Assessment Question Design

Scenario-based assessments have been applied to measuring skills in varied domains, such as writing (Zhang et al. 2019), mathematics (Cayton-Hodges et al. 2012) and science (Bergner and von Davier 2019). Such situated assessment experience can engage learners and facilitate deep thinking by presenting realistic, contextual problems that promote critical thinking and knowledge application (Clark 2010; Herrington and Oliver 2000). Furthermore, as our instructional goals were competency-based, scenario-based assessments are suitable to test students' abilities. Multiple Choice Questions (MCQs) enable automatic grading while providing consistent evaluations across diverse learner groups. When carefully crafted, MCQs can target any cognitive level in Bloom's Taxonomy (Amer 2006), making them a versatile and effective component of our assessment approach. Therefore, the assessment consisted of six MCQs. Each question presents three answer choices, with only one being correct. In Study 1, we chose MCQs as the only assessment question type and situated them into different scenarios based on the learning objectives.

4.3 Data Collection and Analysis

In *Study 1*, we collected both students' written prompts and corresponding AI-grading results. To answer RQ1, we extracted students' last written prompt and graded them manually to assess AI auto-grading. After receiving 483 unique student-written prompts, two researchers used the same rubric to grade the students' written prompts (Table 1). They first randomly selected 15% of the prompt data and conducted human grading based on the rubric. After that, the two researchers met to address the conflicts and iteratively refined the detailed grading book. Then they randomly selected another 15% of the prompt data and both two researchers graded it again; the inter-rater reliability for this round reached higher than 0.92 (almost perfect agreement following (McHugh 2012)). Then the two researchers

	General definition	Scenario 1	Scenario 2	Scenario 3
Relevance	The prompt is related to the topic	✓	✓	✓
Clarity of Purpose	The prompt identifies a specific and clear purpose.	✓	✓	✓
Conciseness	The prompt itself is brief and concise.	✓	✓	✓
Background/Context	The prompt explains why this question was asked, such as the background	✓	✓	✓
Request Elaboration or Explanation	The prompt requests elaboration, extension, or explanation except for seeking a direct response.		✓	✓
Not Explicitly Seeking a Direct Response	The prompt does not ask for a solution to solve a problem or the answers (what are x and y).			✓

Table 1: Grading dimensions and associated tasks for student-written prompts. "✓" means the LLM assesses if the student's prompt meets the definition with a binary response (Yes or No) and then gives corresponding grading explanations.

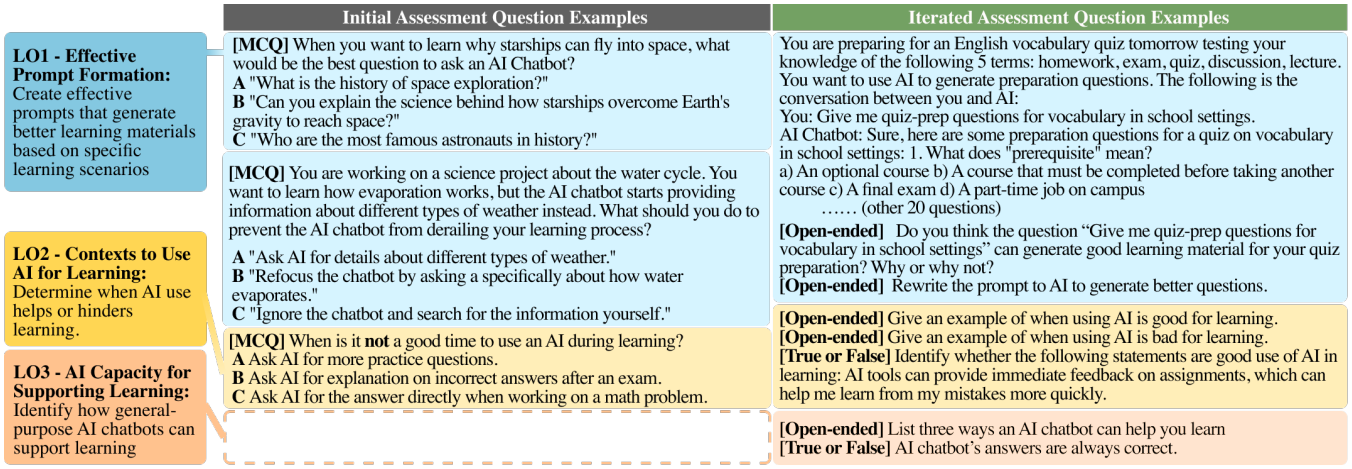

Figure 2: Initial and Iterated Assessment Questions with High-Level Learning Objectives (LO)

met again to resolve the conflicts and graded the rest. We used human grading results for RQ2. For RQ3, we conducted a thematic analysis on student answers. One hundred and thirty-one students finished answering the self-reflection questions. Two researchers conducted a coding book iteratively to analyze their answers and reported the main themes.

4.4 RQ1: What is the accuracy of AI-based grading for students' written prompts?

The grading system is able to grade and provide grading rationale with good quality in most categories. Using the human labels as ground truth, in general, the AI auto-grader achieved an average of 0.92 accuracy when assessing students' prompts and generated high-quality, detailed feedback. The evaluation result is in Table 2.

The auto-grader achieved a high accuracy (higher than 0.9) in *Relevance*, *Background*, and *Elaboration* dimensions. For the inaccurate cases in these dimensions, one pattern is that the auto-grader tended to weigh heavily on some keywords but ignored other keywords (Nguyen et al. 2023). For instance, one prompt was *"You are preparing for a geography exam and have just learned the basic concepts of the water cycle Answer in Sanskrit."* The auto-grader gave it a False in *Relevance* and explained as *"The core of the problem lies in language choice and has nothing to do with the specific content of the water cycle."*.

The accuracy for *Conciseness* and *No Direct Answer* was 0.93 and 0.88. The conflicted cases for *Conciseness* mostly included two types. First, the auto-grader focused on spelling errors and therefore graded them as False, while human graders were more tolerant about this. Second, prompts containing background information might be seen as non-concise by the auto-grader. For example, one student-written prompt was *"Can you tell me in detail about your extracurricular knowledge about cells besides the definition of cells?"*. The auto-grader graded *Conciseness* as False and explained it as *"The question is relatively long and the sentences are not concise enough. It could be more refined."*

As for *No Direct Answer*, sometimes the student asked a general question about *"How to solve a two-variable linear equation?"*, but the auto-grader treated this as the student tried to ask for a direct answer *indirectly* and labeled it as False. On the other hand, sometimes the student asked for *"list the steps to solve this equation 10x+4y=3,-2x+10y=4"*, but auto-grader thought this prompt *"did not directly ask about the values of x and y, but rather asks for a solution."* and gave it True in *No Direct Answer* instead.

Auto-grader received the lowest accuracy in *Purpose* dimension (0.85). The main reason for false positive cases in this dimension is that the auto-grader might over-generate the keywords in a student-written prompt, and consider that as providing a clear purpose. For instance, it counted *"I am*

	Relevance	Purpose	Conciseness	Background	Elaboration	No Answer	Overall
Accuracy of Pass/Fail Classification (1 / 0)	0.98	0.85	0.93	0.96	0.90	0.88	-
Explanation Accuracy (1 / 0.5 / 0)	0.98	0.87	0.96	0.95	0.72	0.91	0.95

Table 2: Grading Accuracy of the Auto-Grader

a junior high school student. Please help me solve this problem using junior high school knowledge." as True in *Purpose* dimension and explained it as *"the student's goal is to solve this problem."* However, simply asking for "solving this problem" without specifying the type or domain does not align with our rubric (*"specific and clear purpose, e.g. explain the math concepts involved"*).

For true negative cases in *Purpose*, sometimes the auto-grader mixed the criteria of *Purpose* and *No Answer*, and therefore considered those involving asking for an answer as False. Another reason is that auto-grader sometimes requires too detailed purpose, which is outside of our targeted scope. For example, one explanation to an AI-graded False case (*"I want to know about DNA"*) is *"Although DNA is mentioned in this question, the purpose is not clear. Students should be more specific about what they want to learn about DNA."* However, as the scenario is novice learners who only finished the first biology class about cells, the auto-grader should not expect the prompt purpose to be that detailed.

4.5 RQ2: How does this activity influence students' prompting ability and confidence in using AI for future learning?

Students answered four 5-scale Likert questions as a pre-survey before doing the pre-test. After finishing the learning activities, they answered the last question in the pre-survey again. Specific questions and the student response distribution are shown in Figure 3.

Students improved at embedding background in prompts with practice. Given that students' scores for each prompt dimension are dichotomous categorical data (Yes-1/No-0) and we wanted to understand how students improved from Q1 to Q3, we conducted a series of McNemar tests by comparing the common four prompting dimensions across all three practice questions (*Relevance, Conciseness, Background, and Purpose*). We found no significant differences in *Relevance* dimension ($p > .90$), *Conciseness* dimension ($p = .286$), and *Purpose* dimension ($p = .617$). Students generally performed well in these dimensions even in the first question. However, most of the students lacked the awareness of embedding background and context information in their prompt at the first question (Q1). For *Background*, as students received feedback and gained more practice opportunities, the proportion of students who did *Background* dimension correctly is significantly higher in Q3 as compared to Q1 ($p = .039$). Additionally, we conducted a Pearson correlation between students' performance in the first practice task and their self-reported prior AI usage frequency. We found that students' frequency of using AI is positively correlated with their average score on the first question ($r = 0.27$, $p = .017$), indicating that students' prior access to AI might influence their initial prompt quality.

Students are more confident in using AI to help learning after the learning activity. After the activity, students' self-reported confidence levels increased by 10.4% on average ($SD = 0.92$), and the Wilcoxon test result indicates the significance of such increment in students' confidence level ($p < .001$). Such improvement can also be observed from Figure 3, the significant rightward shift in the distribution of confidence scores indicates increased confidence in appropriate AI usage in learning.

Lessons Learned: Students understood prompt basics, but their ability to create one varies. Ninety-eight students completed both the pre- and post-test. However, no significant increment ($p = .377$) was found in the post-test (*Mean* = 4.4, *Median* = 5.0, $SD = 0.96$, full score = 6.0) due to the ceiling effect in the pre-test (*Mean* = 4.4, *Median* = 5.0, $SD = 1.04$). In other words, students had the basic skills to identify the prompt quality from a conceptual standpoint, although their prompt writing skills remain varied.

4.6 RQ3: How do students perceive the benefits and challenges of this experience?

We also examined student perceptions including what they learned from this activity, the interests and challenges encountered, and suggestions to refine this activity in the future. From survey responses from *Study 1*, we found the following themes about students' learning experiences.

What do you learn from this learning activity? Most learners (87%) reported that they learned AI-related knowledge, including ***how to use AI in learning*** (e.g. *"I can't ask AI to write answers directly, otherwise I won't learn anything."*), ***how to ask AI questions effectively*** (e.g. *"Adding background and context information in the prompt can help AI to form better answers"*), and ***AI's capabilities*** (e.g. *"AI can help on exam preparation", "AI also has problems that it doesnt́ know about"*). These indicated learners' gain on conceptual knowledge and skills of using AI in general and in learning, aligning with our instructional goals.

What do you like about this learning activity? Fifty-five students showed great interests in ***interaction with AI***. More specifically, students appreciated the chances to ask questions, get responses and evaluate responses with newly learned prompting skills (e.g. *"I am interested in the different responses from the different question-asking approaches"*). The ***learning design*** was also favored by students. Many of them gained great interest in the learning objectives: *"Knowing AI can be used to study or review coursework", "Great to know that AI can not only provide knowledge, but also create exercise questions."* Some

		Relevance	Purpose	Conciseness	Background	Elaboration	No Answer	Overall
Student Final Score (*M*)	Q1	0.89	0.72	0.95	0.04	-	-	-
(Applied human-graded	Q2	0.93	0.65	0.94	0.12	0.59	-	-
scores, Yes-1/No-0)	Q3	0.89	0.75	0.95	0.11	0.52	0.64	-
	Overall	0.90	0.71	0.95	0.09	0.56	0.64	-

Table 3: Student-written prompt scores by dimension. "-" indicates this dimension is not applicable for that question.

students liked the in-time, comprehensive feedback design: *"(the platform) will give detailed instructions after checking the answers I reply."* The scenario design was also favored by 8 students (e.g. *"(I like) the scientific report of water cycle and want to know more about it"*). Additionally, some students favored the ***visual elements*** of the learning platform, including the cartoon logos and its animations (e.g. *"I like the cute cartoon robot holding a sign to tell me whether my answer is correct"*). In short, the parts that students favored highly overlap with our instructional goals.

What is the most challenging part of this learning activity? Students' main challenges could be categorized into productive struggles (Warshauer 2015) which are essential for learning, and elements that increase their extraneous load that should be minimized. The most common productive struggle was ***experiencing difficulties in asking AI questions***, which was the core skill we intended to teach in this activity. While students acknowledged their lack of competency in *"writing a prompt that leads to a more helpful answer"*, many of their responses directly quoted some of the prompting guidelines in our instructional materials, which were strong indicators of students' mastery of tips for writing better prompts. For instance, students admitted the difficulties in writing prompts by saying *"I didn't write a concise question", "I didn't provide context information"*. These precise quotes from the instruction content in survey responses indicated students' knowledge retention of prompting strategies, which served as evidence of our effective activity. Reasons for increasing extraneous load included ***limitations of AI and learning platform*** (e.g. *"It takes so long for AI to response", "I have trouble in logging in"*), ***lack of variation on scenario content*** (e.g. "All scenarios are in science subjects which I'm not interested in"), and noticeably, ***limitations of students' basic computer skills*** (N = 22, e.g. *"I am not good at typing, it is tedious to type a lot to ask AI questions"*). Some students struggled in typing, copying, and pasting information, which might hinder their prompt construction, even if they understood how to improve one.

5 Study 2: Assessment Iteration Follow-Up

Results from Study 1 revealed a ceiling effect in the pre-test, indicating that students in our study had a basic understanding to *identify* learning-oriented prompts. However, practice performance showed that they had not yet mastered *writing* such prompts effectively (Table 3). Therefore, to better measure their prompting literacy learning (Judson 2012; Staus, O'Connell, and Storksdieck 2021), we iterated the assessment and conducted *Study 2* to evaluate the assessment quality. Study 2 utilized the same procedure and materials with a similar population, except for the iterated assessment questions. Study 2 focused on RQ4 about assessment questions:

RQ4: Does the revised assessment demonstrate improved quality in measuring students' prompting literacy?

5.1 Iterated Assessment Design

The iterated assessment included 15 questions: ten True or False (TF) questions and five Open-Ended (OE) questions. Specifically, TF1 to TF6 and OE1 targeted at LO-AI Capacity for Supporting Learning; TF7 to TF10 and OE2, OE3 targeted at LO-Contexts to Use AI for Learning; and OE4 and OE5 targeted at LO-Effective Prompt Formation.

Replace multiple-choice questions with True / False and open-ended questions The lack of information about learners' prior knowledge can reduce the effectiveness of MCQ distractors. In contrast, TF questions require students to make a decision on each item individually, offering more data points to assess both the learners' true knowledge level and the quality of the question items, particularly at lower granularity (Brassil and Couch 2019). In addition, to overcome difficulties in forming high-quality distractors in MCQs, one approach is to collect learners' common misconceptions as MCQ items. OE1, OE2, and OE3 are designed as a misconception collectors for *LO-AI Capacity for Supporting Learning* and *LO2-Contexts to Use AI for Learning* to facilitate further iterations on assessment questions.

The gap between high MCQ scores and poor prompt-writing performance from Study 1 suggested that *LO-Effective Prompt Formation* involves higher-order cognitive skills (falling under the *analyze* and *create* levels of Bloom's Revised Taxonomy (Krathwohl 2002)), and developing MCQs that effectively assess these higher cognitive levels is challenging even for expert instructors in traditional disciplines (Douglas, Wilson, and Ennis 2012), let alone for an emerging topic. Therefore, we switched to open-ended questions and designed OE4 (*Given a prompt and the AI Chatbot response, do you think the question can generate good learning material for your quiz preparation? Why or why not?*) to assess students' analyzing skill, and OE5 (*Rewrite the question to AI to generate better preparation questions*) to assess prompt-writing skill more directly.

Add abstract-level questions to increase the question difficulty variations To more accurately estimate learners' prior knowledge, assessment questions should be designed with varying levels of difficulty and knowledge depth under the same learning objective. As such, abstract-level questions are included alongside concrete, scenario-based questions to assess learners across a broader spectrum of difficulty within the same objective.

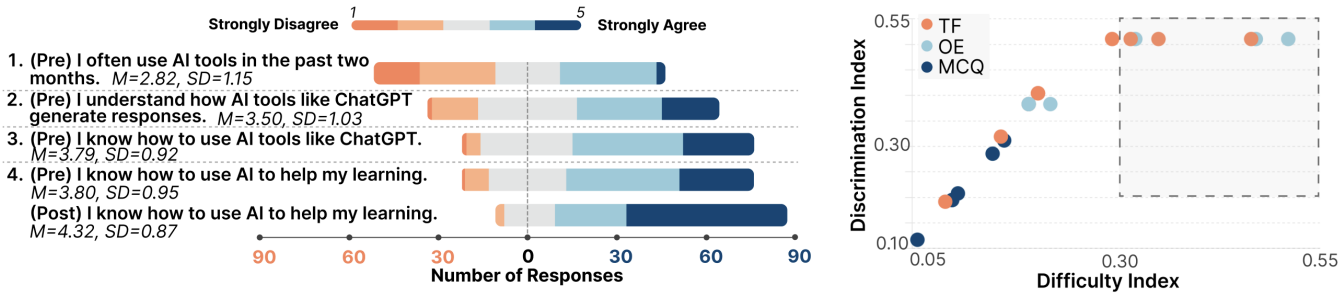

Figure 3: Left: Student responses to likert questions in the pre-test and post-test; Right: difficulty level vs discrimination index for all questions in both assessments. True or False (TF) questions and Open-Ended (OE) questions are in the revised assessment, and Multiple Choice Questions (MCQ) are in the original assessment.

5.2 RQ4: Does the revised assessment better evaluate the targeted learning objectives?

We evaluated the assessment question quality using difficulty and discrimination indices, and reported the overall reliability using Cronbach's Alpha. The comparison of the initial and revised assessment is used to answer RQ4. In particular, we compared the quality of these two assessments using the data collected from two studies and three widely-used assessment evaluation matrices: difficulty, discrimination, and reliability (Nitko 1996; Salkind 2017). Our analyses yielded two preliminary findings:

Item difficulty and discrimination are improved in the iterated version. The difficulty level and discrimination index are calculated based on the formula definitions provided in the previous work (Boopathiraj and Chellamani 2013). Figure 3 displays the difficulty level and discrimination index for all questions. A good discrimination index should be no less than 0.2 and ideally approach 1, while a good difficulty level falls within the range of [0.3, 0.7]. None of the MCQ questions fall into the desired range, while 60% of OE and 30% of TF questions satisfied the requirement, indicating an improvement on the item quality as assessment .

Question reliability should be evaluated with a larger population later With the initial success on difficulty level and discrimination index, Cronbach's Alpha analysis showed moderate internal reliability for both assessment versions: 0.68 for the original and 0.58 for the iterated version. Although slightly below the 0.70 benchmark, these values are reasonable given the small sample size and limited number of items. The lower reliability in the iterated version may reflect greater item diversity or more targeted revisions. Larger-scale administrations will be needed to confirm these patterns and further strengthen internal consistency.

6 Lessons Learned and Future Work

In this work, we first designed and implemented a K-12 prompting literacy learning module featuring scenario-based chatbot interactions and AI-driven auto-assessment, and then conducted two classroom studies: *Study 1* to measure technical capabilities and learning experience, and *Study 2*, which built on data-informed refinements from *Study 1* to evaluate the assessment quality.

First, our results provided ***design implications of teaching prompting literacy and AI***. By experiencing the learning-by-doing prompting activities with LLM-generated immediate elaborated feedback in various scenarios, students gained more experience in including contexts in their prompts and increased confidence in prompt writing for learning. However, the lack of improvement in other aspects (e.g., conciseness, elaboration) and post-test performance may be attributed to dosage effects (Zhai et al. 2010), suggesting that more exposure is needed for aspects that students are less familiar with. Survey responses showed that task scenarios increased students' interest and motivation, but students also wanted more relatable and diverse topics to match their varied interests. We also note that our study did not compare this approach to other AI literacy instructional methods; future work should conduct comparative studies or use this study as a benchmark for further iterations.

Secondly, this work introduced ***a scalable learning and assessment platform***. The AI-based auto-grader achieved high accuracy in most dimensions, indicating its potential to provide immediate, detailed feedback for learning prompt writing. However, certain limitations remain in the *No Direct Answer* and *Purpose* dimensions, suggesting future refinement. The AI's tendency to misinterpret the prompt intention or overly focus on language mechanics highlights where human oversight and rubric refinement improve reliability.

Lastly, this work contributed to the development of learning and assessment materials for prompting literacy, informed by iterative design and data-driven refinement. As the first local study on prompting literacy, students' baseline knowledge was unclear, making classroom-based iteration essential for ensuring instructional effectiveness. Increasing the variety and number of assessment items will help better surface learners' misconceptions and knowledge gaps. These initial efforts establish a foundation for prompting literacy instruction. Future work should scale assessments to larger populations and items, using models such as Item Response Theory (IRT) (Embretson and Reise 2000) to produce statistically robust measures of assessment quality.

References

Amer, A. 2006. Reflections on Bloom's revised taxonomy. *Electronic Journal of Research in Educational Psychology*,

4(1): 213–230.

Bergner, Y.; and von Davier, A. A. 2019. Process data in NAEP: Past, present, and future. *Journal of Educational and Behavioral Statistics*, 44(6): 706–732.

Bonwell, C. C.; and Eison, J. A. 1991. *Active learning: Creating excitement in the classroom. 1991 ASHE-ERIC higher education reports.* ERIC.

Boopathiraj, C.; and Chellamani, K. 2013. Analysis of test items on difficulty level and discrimination index in the test for research in education. *International journal of social science & interdisciplinary research*, 2(2): 189–193.

Brassil, C. E.; and Couch, B. A. 2019. Multiple-true-false questions reveal more thoroughly the complexity of student thinking than multiple-choice questions: a Bayesian item response model comparison. *International Journal of STEM Education*, 6(1): 1–17.

Cayton-Hodges, G. A.; Marquez, E.; Keehner, M.; Laitusis, C.; van Rijn, P.; Zapata-Rivera, D.; and Hakkinen, M. 2012. Technology enhanced assessments in mathematics and beyond: Strengths, challenges, and future directions. In *ETS Invitational Research Symposium on Technology Enhanced Assessments, Washington, DC*.

Clark, I. 2010. Formative assessment:'There is nothing so practical as a good theory'. *Australian Journal of Education*, 54(3): 341–352.

Dennison, D. V.; Garcia, R. C.; Sarin, P.; Wolf, J.; Bywater, C.; Xie, B.; and Lee, V. R. 2024. From consumers to critical users: Prompty, an AI literacy tool for high school students. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 23300–23308.

Denny, P.; Kumar, V.; and Giacaman, N. 2023. Conversing with copilot: Exploring prompt engineering for solving cs1 problems using natural language. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, 1136–1142.

Denny, P.; Leinonen, J.; Prather, J.; Luxton-Reilly, A.; Amarouche, T.; Becker, B. A.; and Reeves, B. N. 2023. Promptly: Using prompt problems to teach learners how to effectively utilize ai code generators. *arXiv preprint arXiv:2307.16364*.

Douglas, M.; Wilson, J.; and Ennis, S. 2012. Multiple-choice question tests: a convenient, flexible and effective learning tool? A case study. *Innovations in Education and Teaching International*, 49(2): 111–121.

Ekin, S. 2023. Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices. *Authorea Preprints*.

Embretson, S. E.; and Reise, S. P. 2000. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates.

Giray, L. 2023. Prompt engineering with ChatGPT: a guide for academic writers. *Annals of biomedical engineering*, 51(12): 2629–2633.

Henkel, O.; Hills, L.; Boxer, A.; Roberts, B.; and Levonian, Z. 2024. Can Large Language Models Make the Grade? An Empirical Study Evaluating LLMs Ability To Mark Short Answer Questions in K-12 Education. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, 300–304.

Herrington, J.; and Oliver, R. 2000. An instructional design framework for authentic learning environments. *Educational technology research and development*, 48(3): 23–48.

Hou, X.; Xiao, R.; Ye, R.; Liut, M.; and Stamper, J. 2025. Exploring Student Choice and the Use of Multimodal Generative AI in Programming Learning. *arXiv preprint arXiv:2510.05417*.

Hwang, Y.; Lee, J. H.; and 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*.

Judson, E. 2012. Learning about bones at a science museum: examining the alternate hypotheses of ceiling effect and prior knowledge. *Instructional Science*, 40: 957–973.

Kazemitabaar, M.; Hou, X.; Henley, A.; Ericson, B. J.; Weintrop, D.; and Grossman, T. 2023. How novices use LLM-based code generators to solve CS1 coding tasks in a self-paced learning environment. In *Proceedings of the 23rd Koli calling international conference on computing education research*, 1–12.

Koedinger, K. R.; Kim, J.; Jia, J. Z.; McLaughlin, E. A.; and Bier, N. L. 2015. Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In *Proceedings of the second (2015) ACM conference on learning@ scale*, 111–120.

Korzynski, P.; Mazurek, G.; Krzypkowska, P.; and 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.

Krathwohl, D. R. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice*, 41(4): 212–218.

Lee, V. R.; Pope, D.; Miles, S.; and Zárate, R. C. 2024. Cheating in the age of generative AI: A high school survey study of cheating behaviors before and after the release of ChatGPT. *Computers and Education: Artificial Intelligence*, 7: 100253.

Li, H.; Xiao, R.; Nieu, H.; Tseng, Y.-J.; and Liao, G. 2025. "From Unseen Needs to Classroom Solutions": Exploring AI Literacy Challenges & Opportunities with Project-Based Learning Toolkit in K-12 Education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 29145–29152.

Lo, L. S. 2023. The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship*, 49(4): 102720.

Lozano, A.; and Blanco Fontao, C. 2023. Is the education system prepared for the irruption of artificial intelligence? A study on the perceptions of students of primary education degree from a dual perspective: Current pupils and future teachers. *Education Sciences*, 13(7): 733.

Maloy, R. W.; and Gattupalli, S. 2024. Prompt literacy. *EdTechnica: The Open Encyclopedia of Educational Technology*.

McHugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3): 276–282.

Meskó, B. 2023. Prompt engineering as an important emerging skill for medical professionals: tutorial. *Journal of medical Internet research*, 25: e50638.

Nayak, J.; Keane, T.; Linden, T.; and Molnar, A. 2023. Teaching high school students artificial intelligence by programming Chatbots. In *Teaching Coding in K-12 Schools: Research and Application*, 263–276. Springer.

Ng, D. T. K.; Leung, J. K. L.; Su, M. J.; Yim, I. H. Y.; Qiao, M. S.; and Chu, S. K. W. 2023. *AI literacy in K-16 classrooms*. Springer International Publishing AG.

Nguyen, H. A.; Stec, H.; Hou, X.; Di, S.; and McLaren, B. M. 2023. Evaluating chatgpt's decimal skills and feedback generation in a digital learning game. In *European Conference on Technology Enhanced Learning*, 278–293. Springer.

Nitko, A. J. 1996. *Educational assessment of students*. ERIC.

Salkind, N. J. 2017. *Tests & measurement for people who (think they) hate tests & measurement*. Sage Publications.

Schank, R. C.; Berman, T. R.; and Macpherson, K. A. 2013. Learning by doing. In *Instructional-design theories and models*, 161–181. Routledge.

Staus, N. L.; O'Connell, K.; and Storksdieck, M. 2021. Addressing the ceiling effect when assessing STEM out-of-school time experiences. In *Frontiers in education*, volume 6, 690431. Frontiers Media SA.

Touretzky, D.; Martin, F.; Seehorn, D.; Breazeal, C.; and Posner, T. 2019. Special session: AI for K-12 guidelines initiative. In *Proceedings of the 50th ACM technical symposium on computer science education*, 492–493.

Trucano, M. 2023. AI and the next digital divide in education.

Tseng, Y.-J.; Xiao, R.; Bogart, C.; Savelka, J.; and Sakr, M. 2024. Assessing the Efficacy of Goal-Based Scenarios in Scaling AI Literacy for Non-Technical Learners. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2*, 1842–1843.

Uanachain, D. M. N.; and Aouad, L. I. 2025. Generative AI in Education: Rethinking Learning, Assessment & Student Agency for the AI Era. *Thresholds in Education*, 48(1): 111–132.

Van Brummelen, J.; Shen, J. H.; and Patton, E. W. 2019. The popstar, the poet, and the grinch: Relating artificial intelligence to the computational thinking framework with block-based coding. In *Proceedings of International Conference on Computational Thinking Education*, volume 3, 160–161.

Vasilyeva, E.; De Bra, P.; and Pechenizkiy, M. 2008. Immediate elaborated feedback personalization in online assessment. In *European Conference on Technology Enhanced Learning*, 449–460. Springer.

Wang, M.; Wang, M.; Xu, X.; Yang, L.; Cai, D.; and Yin, M. 2023. Unleashing ChatGPT's power: A case study on optimizing information retrieval in flipped classrooms via prompt engineering. *IEEE Transactions on Learning Technologies*.

Wang, Z.; Gong, S.-Y.; Xu, S.; and Hu, X.-E. 2019. Elaborated feedback and learning: Examining cognitive and motivational influences. *Computers & Education*, 136: 130–140.

Warshauer, H. K. 2015. Productive struggle in middle school mathematics classrooms. *Journal of Mathematics Teacher Education*, 18: 375–400.

White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; and Schmidt, D. C. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Williamson, B.; Molnar, A.; and Boninger, F. 2024. Time for a Pause: Without Effective Public Oversight, AI in Schools Will Do More Harm Than Good. *Commercialism in Education Research Unit*.

Woo, D. J.; Guo, K.; and Susanto, H. 2024. Exploring EFL students' prompt engineering in human–AI story writing: an activity theory perspective. *Interactive Learning Environments*, 1–20.

Xiao, R.; Hou, X.; and Stamper, J. 2024. Exploring How Multiple Levels of GPT-Generated Programming Hints Support or Disappoint Novices. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–10.

Xiao, R.; Hou, X.; Ye, R.; Kazemitabaar, M.; Diana, N.; Liut, M.; and Stamper, J. 2025a. Improving Student-AI Interaction Through Pedagogical Prompting: An Example in Computer Science Education. *arXiv preprint arXiv:2506.19107*.

Xiao, R.; Xiao, Q.; Hou, X.; Moletsane, P. P.; Li, H. J.; Shen, H.; and Stamper, J. 2025b. Do Teachers Dream of GenAI Widening Educational (In) equality? Envisioning the Future of K-12 GenAI Education from Global Teachers' Perspectives. *arXiv preprint arXiv:2509.10782*.

Yim, I. H. Y.; and Su, J. 2024. Artificial intelligence (AI) learning tools in K-12 education: A scoping review. *Journal of Computers in Education*, 1–39.

Zamfirescu-Pereira, J.; Wong, R. Y.; Hartmann, B.; and Yang, Q. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21.

Zhai, F.; Raver, C. C.; Jones, S. M.; Li-Grining, C. P.; Pressler, E.; and Gao, Q. 2010. Dosage effects on school readiness: Evidence from a randomized classroom-based intervention. *Social Service Review*, 84(4): 615–655.

Zhang, M.; van Rijn, P. W.; Deane, P.; and Bennett, R. E. 2019. Scenario-based assessments in writing: An experimental study. *Educational Assessment*, 24(2): 73–90.

Zhang, P.; and Tur, G. 2024. A systematic review of ChatGPT use in K-12 education. *European Journal of Education*, 59(2): e12599.