

Reflection-Satisfaction Tradeoff: Investigating Impact of Reflection on Student Engagement with AI-Generated Programming Hints

Heeryung Choi^a, Tung Phung^b, Mengyan Wu^a, Adish Singla^b, Christopher Brooks^a

^a*University of Michigan, Ann Arbor, USA*

^b*Max Planck Institute for Software Systems, Saarbrücken, Germany*

Abstract

Generative AI tools, such as AI-generated hints, are increasingly integrated into programming education to offer timely, personalized support. However, little is known about how to effectively leverage these hints while ensuring autonomous and meaningful learning. One promising approach involves pairing AI-generated hints with reflection prompts, asking students to review and analyze their learning, when they request hints. This study investigates the interplay between AI-generated hints and different designs of reflection prompts in an online introductory programming course. We conducted a two-trial field experiment. In Trial 1, students were randomly assigned to receive prompts either *before* or *after* receiving hints, or *no prompt* at all. Each prompt also targeted one of three SRL phases: *planning*, *monitoring*, and *evaluation*. In Trial 2, we examined two types of prompt guidance: *directed* (offering more explicit and structured guidance) and *open* (offering more general and less constrained guidance). Findings show that students in the *before*-hint (RQ1), *planning* (RQ2), and *directed* (RQ3) prompt groups produced higher-quality reflections but reported lower satisfaction with AI-generated hints than those in other conditions. Immediate performance did not differ across conditions. This negative relationship between reflection quality and hint satisfaction aligns with previous work on student mental

Email addresses: heeryung@umich.edu (Heeryung Choi), mphung@mpi-sws.org (Tung Phung), mengyanw@umich.edu (Mengyan Wu), adishs@mpi-sws.org (Adish Singla), brooks@umich.edu (Christopher Brooks)

effort and satisfaction. Our results highlight the need to reconsider how AI models are trained and evaluated for education, as prioritizing user satisfaction can undermine deeper learning.

Keywords:

AI-generated hints, Reflection, Self-regulated learning, Generative AI, Reflection-Satisfaction Tradeoff

1. Introduction

Generative AI systems, such as ChatGPT (OpenAI, 2023), are rapidly transforming programming courses. These tools are increasingly employed to enhance educational experiences by offering on-demand instant assistance tailored to students’ needs (Denny et al., 2024a; Phung et al., 2023; Finnie-Ansley et al., 2022). Among the various AI-powered educational applications, AI-generated hints have drawn particular attention. AI-generated hints are designed to guide students through complex tasks by providing timely, context-sensitive feedback, which may significantly affect students’ problem-solving and help-seeking. Recent studies have indicated teachers’ and students’ positive reception to using this tool in educational contexts (Pankiewicz and Baker, 2024; Burner et al., 2025).

However, research has found mixed outcomes regarding the educational benefits of integrating generative AI and AI-generated hints. Some studies reported a negligible or negative effect of students’ AI use on their performance, such as a decrease in deep learning (Zamfirescu-Pereira et al., 2025; Bastani et al., 2024). In another work by Pankiewicz and Baker (2023), the positive impact of AI-generated hints on students’ performance did not last once these hints were removed.

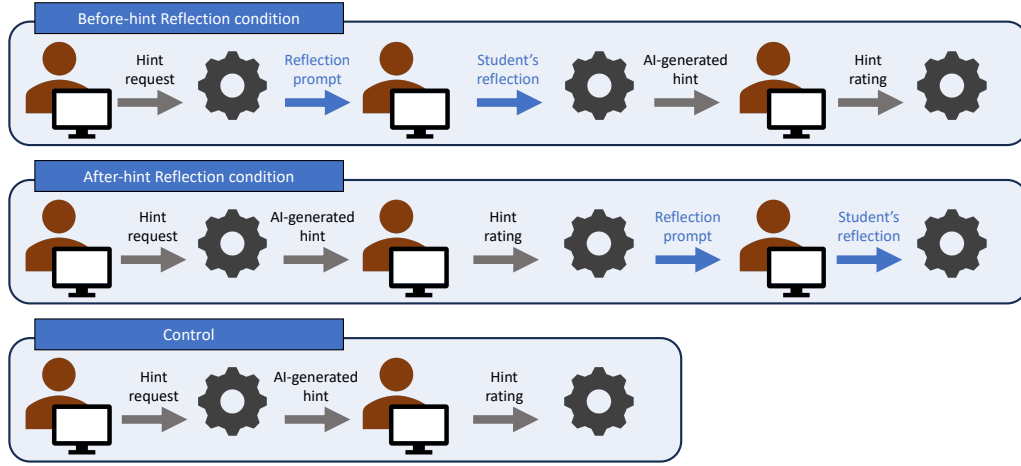
Several studies attributed the discouraging impact of AI on education to *metacognitive laziness*, the tendency of students to avoid and disengage from effortful learning processes (Fan et al., 2025). As students exhibit metacognitive laziness and increasingly turn to AI support, they decrease opportunities to develop their fundamental learning skills, such as self-regulated learning (SRL) and critical thinking (Gerlich, 2025; Lee et al., 2025; Zhan and Yan, 2025). This concern is further underscored by the recent Anthropic Education Report, which showed that approximately 70% of student-AI interaction was to delegate higher-order thinking to AI agents (Handa et al., 2025). These findings highlight the pressing need for further research into instruc-

tional design and guidance of integrating AI-generated hints to address the issue of metacognitive laziness and more effectively leverage the educational benefits of AI (Phung et al., 2023; Denny et al., 2024b,a).

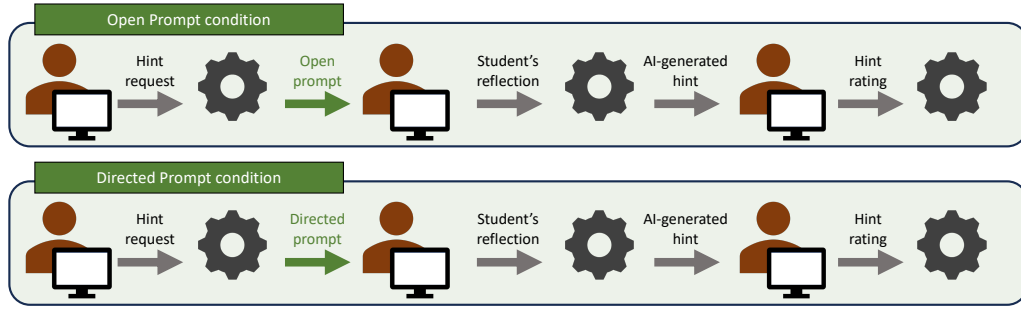
One promising solution could be reflection-focused interventions. Reflective practices encourage students to critically analyze their thinking, problem-solving strategies, and errors, thereby fostering self-awareness and iterative improvement (Zimmerman, 2002; Tankelevitch et al., 2024; Choi et al., 2023). These benefits of reflective practices have been well studied and are frequently used as a scaffold to encourage SRL and critical thinking across educational contexts, including environments that provide students with hints. Previous research has demonstrated that prompting reflection is specifically effective in addressing students’ metacognitive laziness during their interactions with predefined, less personalized hints (Choi et al., 2023; Wang et al., 2017; Davis and Linn, 2000). When students were prompted to reflect, they seemed to more actively and meaningfully engage with hints; they showed higher self-efficacy, and better immediate and delayed performances (Choi et al., 2023; Bardach et al., 2021). Compared to simply preventing unproductive learning behaviors, reflection is often a more effective approach because it not only limits undesirable hint use but also fosters active engagement with learning that is likely to develop into long-term learning behaviors (Aleven et al., 2006; Muñoz-Merino et al., 2013).

Given these positive impacts of reflection practices, prompting for reflection seemed to be a promising approach to address metacognitive laziness and enhance students’ productive engagements with the AI-generated hints. However, the specific mechanism of how reflection interacts with AI-generated hints in learning environments remains underexplored (Margulieux et al., 2024; Marwan et al., 2019; Prasad and Sane, 2024; Kumar et al., 2024; Singh et al., 2024). Understanding the underlying mechanism is crucial to ensure reflection intervention is effectively implemented alongside AI-generated hints in real-world classrooms.

To address this gap, we conducted two field experiments to examine the impact of reflection practices on student metacognitive laziness during interactions with AI-generated hints in an online introductory data science programming course (see Figure 1). The study was conducted using a system that provided AI-generated hints on demand within students’ assignment environment (JupyterLab (Team, 2018)). The first of these trials focused on understanding the effect of the placement and targeted SRL phases of reflection prompts in the help-seeking process. The second trial focused on



(a) Student interactions with the support system in Trial 1.



(b) Student interactions with the support system in Trial 2.

Figure 1: Illustration of the two randomized controlled trials involved in our study. Subfigure (a) shows student interactions with the support system in Trial 1. Students were randomly assigned to one of three conditions. The Before-hint and After-hint Reflection conditions were prompted for reflection, differing in timing relative to receiving hints. The Control condition only receives hints without reflection prompts. In the Before-hint and After-hint Reflection condition, each time a student requested a hint, one of the three prompts targeting different SRL phases (planning, monitoring, or evaluation) was randomly selected to show to them. See Table 1 for all prompts and Table 7 for more examples of students' reflections corresponding to different prompts in Trial 1. Subfigure (b) shows student interactions with the support system in Trial 2. Students were randomly assigned to one of two conditions: Open Prompt and Directed Prompt. Depending on the condition, a student was always shown one corresponding reflection prompt whenever they requested a hint. See Table 2 for the two reflection prompts and Table 8 for more examples of students' reflections in Trial 2.

understanding how the amount of guidance in reflection prompts affects student behaviors and satisfaction with provided hints. Taken together, these trials address the following three research questions:

- **RQ1:** How does the placement of reflection prompts (*before* vs. *after* receiving an AI-generated hint) affect students' hint-seeking experiences, problem-solving performance, and reflective responses?
- **RQ2:** How does the SRL phase emphasized in reflection prompts (*planning*, *monitoring*, and *evaluation*) affect students' hint-seeking experiences, problem-solving performance, and reflective responses?
- **RQ3:** How does the amount of guidance in reflection prompt (*directed*, providing specific guidance on what to address in a reflection, vs. *open*, where minimal guidance is provided) affect students' hint-seeking experiences, problem-solving performance, and reflective responses?

Through these trials, we discovered several important relationships between hint-seeking behaviors, reflective writing, and AI hint generation. Firstly, across trials, we found a consistent inverse relationship between student satisfaction with AI-generated hints and the quality of the students' reflective responses. In addition to that general trend, students who received reflection prompts *before* receiving AI-generated hints produced higher-quality reflections but reported lower satisfaction with the hints compared to those who were prompted *after* hints or not prompted (RQ1). *Planning*-focused reflection prompts generally led to higher-quality reflections and lower satisfaction than *monitoring*- or *evaluation*-focused prompts (RQ2). *Directed* prompts yielded higher-quality reflections and lower hint satisfaction than *open* prompts (RQ3).

The consistent findings of an inverse relationship between satisfaction over AI-generated hints and reflection quality point to a broader pedagogical tension between optimizing AI systems for immediate user satisfaction and the goal of fostering deeper and effortful learning. This tension invites educational stakeholders to reconsider foundational systemic aspects of higher education structures in a time where AI can outperform traditional educational support roles for learners. Specifically, this work contributes to the broader discussions on learning with AI in several ways, as it:

- Demonstrates the tension between AI systems optimization for user satisfaction and the pedagogical goal for effortful learning, highlighting

the need to prioritize pedagogical values over satisfaction in the design and development of educational AI.

- Provides empirical evidence from an authentic learning environment on how reflective scaffolding during students’ interaction with AI-generated hints shapes their hint-seeking experiences, reflection quality, and performance.
- Underscores the importance of purposefully designing AI-powered educational tools to actively cultivate students’ SRL skills, the key to critical thinking.

In summary, this work advances understanding of how reflection and AI-generated hints can be integrated to foster more meaningful learning, and provides a foundation for rethinking how learning systems and AI models might evolve in response to new educational realities.

2. Related Work

2.1. *Learning with Reflection*

Reflection is widely recognized as a process of critically reviewing one’s learning experiences to strengthen one’s comprehension and enrich insights (Dewey, 1997; Boud et al., 1985; Boud, 2013; O’Rourke, 1998). Many studies have demonstrated the metacognitive, cognitive, and emotional benefits of reflection in programming education (Bardach et al., 2021; Wang et al., 2017; Choi et al., 2023). Choi et al. (2023) prompted students to reflect after completing programming assignments to encourage more pedagogically meaningful engagement with assignment hints. The authors found that students who received reflection prompts reported significantly higher perception of learning and performed better on both immediate and delayed post-tests. Kumar et al. (2024) conducted randomized controlled trials to investigate the effect of post-lesson reflection activities guided by a GPT-3 large language model in an undergraduate computer science course. The authors found that the reflection group not only reported an increase in self-confidence but also outperformed the control group on post-tests after two weeks. Lishinski and Yadav (2021b) found that undergraduate students who completed self-evaluation tasks that targeted reflection and monitoring outperformed their control group in computer science project scores.

Historically, researchers have actively explored how to design reflection prompts to enhance the benefits of reflection. In this section, we focus on three key aspects of reflection prompts: the placement of the prompt (*before* and *after*), its alignment with SRL phases (*planning*, *monitoring*, and *evaluation*), and the amount of guidance it provides (*directed* and *open*) (Bardach et al., 2021; Davis and Linn, 2000; Wang et al., 2017; Mann et al., 2009).

2.1.1. Placement

Reflection can take place during any stage of learning and activates different thinking skills. Specifically, reflection before action aims to plan for better future consequences. During this prospective reflection, students typically review relevant prior experience or external resources, identify optimal strategies, and simulate their actions (Edwards, 2017; Schön, 2017; Mann et al., 2009; Taguma and Barrera, 2019).

On the other hand, reflection after action generally centers self-evaluation on strategies used and results achieved, drawing lessons about what was effective and identifying areas for improvement (Taguma and Barrera, 2019). For instance, students may realize that incorporating visualizations was more straightforward than expected, and that it effectively enhanced their project’s quality (Loksa et al., 2020).

2.1.2. Self-Regulated Learning Phases

Reflection is often considered as the final phase of the SRL cycle, where learners review and evaluate their performance after completing a learning task. Zimmerman’s cyclical model describes SRL as unfolding across three interrelated phases: forethought (planning), performance (monitoring), and self-reflection (evaluation). (Zimmerman, 2002). However, reflection is not limited to the final evaluation stage of learning. Reflection can occur throughout all stages of learning (See 2.1.1). When reflection becomes a part of planning, monitoring, and evaluation phases, it helps students foster metacognitive awareness, adaptive learning strategies, and deeper engagement with learning activities. A growing body of research in programming education shows reflection’s value at every stage.

Studies have shown that students who engage in reflection before starting to code (reflection-before-action, reflection-for-action) tended to develop a deeper understanding of core concepts (Edwards, 2004a). For example, when students collaboratively reflected on similar worked examples before their task, they generally learned more conceptual knowledge than those who did

not reflect and immediately began problem-solving (Sankaranarayanan et al., 2022). The study findings also showed that spending more time on reflective planning did not hurt their problem-solving performance.

As a part of the monitoring phase (reflection-in-action), students reflect on their previous practices and test out their hypotheses to decide their next action (Schön, 2017). For instance, Edwards (2004b) designed an intervention that encouraged students to reflect during software testing activities. The author found that those students produced 45% fewer bugs per 1000 lines of code. The intervention also increased their confidence when making changes to their code. Zarestky et al. (2022) demonstrated that students who annotated lecture videos as a reflective activity outperformed their peers who did not participate in annotation.

After task completion, reflective interventions are most often implemented as self-evaluation activities (reflection-on-action) as mentioned in 2.1.1 (Schön, 2017). For example, Lishinski and Yadav (2021a) examined the impact of self-evaluation exercises in an introductory programming course and found that students who participated in these activities earned significantly higher project scores.

2.1.3. Amount of Guidance

The reflection prompt, which directs the learner to the reflective writing task, can provide different degrees of guidance. For instance, open prompts encourage students to reflect in ways that are not tied to specific contexts, allowing for greater autonomy and flexibility. These prompts are often open-ended in nature, such as “As of now, we are thinking of...”, where students are invited to complete the prompt, setting their own direction for the reflection. Open prompts can help to foster personal meaning-making and deeper engagement, especially for learners who are self-motivated or have developed metacognitive skills (Davis, 2003). In contrast, directed prompts provide more explicit guidance, suggesting productive directions based on established principles of cognition and instruction, such as “Looking back on this project, are there specific areas where you wish you had spent more time? If so, which ones?” (Davis and Linn, 2000; Davis, 2003; Rum and Ismail, 2017). Directed prompts act as a stronger scaffold of the reflection activity, supporting students who may be less experienced with reflective practice or who benefit from additional examples when organizing their thoughts (Davis and Linn, 2000; Johnson and Mayer, 2010).

The effectiveness of open versus directed prompts has been the subject

of considerable research and debate. Some studies have found that open prompts can foster more cohesive thought, deeper integration of knowledge, and greater expansion of ideas, particularly among students with higher autonomy (Davis and Linn, 2000). Other research suggests that directed prompts are especially effective when combined with structured activities designed to support project completion or specific learning goals (Johnson and Mayer, 2010).

Taken together, these underexplored questions and mixed findings in the literature highlight the importance of considering the placement, associated SRL phases, and the amount of guidance in the reflection prompt when designing interventions to support effective reflection practices of students.

2.2. AI and AI-generated Hints in Programming Education

The integration of generative AI into programming education has attracted substantial interest, with numerous studies exploring the potential benefits and challenges of this technology (Shen et al., 2024; Phung et al., 2023; Prather et al., 2023; Denny et al., 2024b; Becker et al., 2023; Yilmaz and Yilmaz, 2023). Research on student-AI interaction has found that students are generally receptive to AI support and willing to incorporate it into their studies (Ma et al., 2024; Valova et al., 2024). This widespread interest highlights the importance of a timely investigation into the impacts that AI brings to learning environments.

Existing studies on the impact of AI on learning, however, have shown mixed results. Concerningly, Kosmyna et al. (2025) reported that over a four-month period, subjects who used generative AI for essay writing exhibited a substantial decline in brain connectivity and reduced engagement in alpha and beta networks, relative to peers who did not use AI. Regarding AI-generated hints, Wang et al. (2024) reported an encouraging finding that AI-generated hints reduce students’ repetitions of errors. On the other hand, research from Bastani et al. (2024) and Zamfirescu-Pereira et al. (2025) found no significant learning gains from AI support. Pankiewicz and Baker (2023) noted that when AI-generated hints were no longer available, students who had previously received such hints had lower first-attempt success rates compared to the control group. These findings emphasize the need for further research on mitigating the over-assistance of AI in educational contexts to benefit students’ fundamental learning skills in long term. Addressing this need, our work investigates the use of reflection prompts in a learning environment where AI-generated hints are available, examining the

interplay between these two forms of support and how reflection prompts can be effectively employed in this setting.

2.3. *Interplay of Reflection and AI-generated Hints*

The intersection between AI-generated hints and reflection in education is a rapidly developing area. However, existing work examined only the correlation between AI usage and self-regulation, without investigating how students’ use of AI-generated hints might influence or be influenced by their reflective practices. For instance, Margulieux et al. (Margulieux et al., 2024) explored the relationship between AI usage and students’ self-regulation, finding no correlation between these elements. This leaves open questions about the dynamic relationship between reflection and AI-generated support, such as how engagement with AI-generated hints could shape students’ reflective practices, and vice versa.

To address these gaps, we investigate the influences between reflection and AI-generated hints. Specifically, we examine how the availability of reflection prompts affects students’ AI-generated hint uses and how variations in reflection practice (prompt placement, targeted SRL phases, and amount of guidance in the prompt) influence students’ overall experiences with AI-generated hints, problem-solving performance, and reflection practices.

3. Study Design

This study involved two randomized controlled trials (See Figure 1), both used a custom support system that can deliver reflection prompts and AI-generated hints. The trials were conducted in two separate offerings of the same online programming course. Each trial examined different factors influencing students’ reflection practices, as well as the effects of these practices on learning behaviors.

Trial 1 investigated the impact of reflection prompts (*presence* vs. *absence*), the placement of prompts (*before* vs. *after* a hint), and the SRL phase targeted by the prompt (*planning* vs. *monitoring* vs. *evaluation*) (Zimmerman, 2002) (See Figure 2 for an example). Trial 2 focused on the impact of prompt guidance type (*directed* vs. *open*).

3.1. *Course Context*

This study was conducted in two runs of the same introductory data science course using Python, which took place in May 2024 and August 2024.

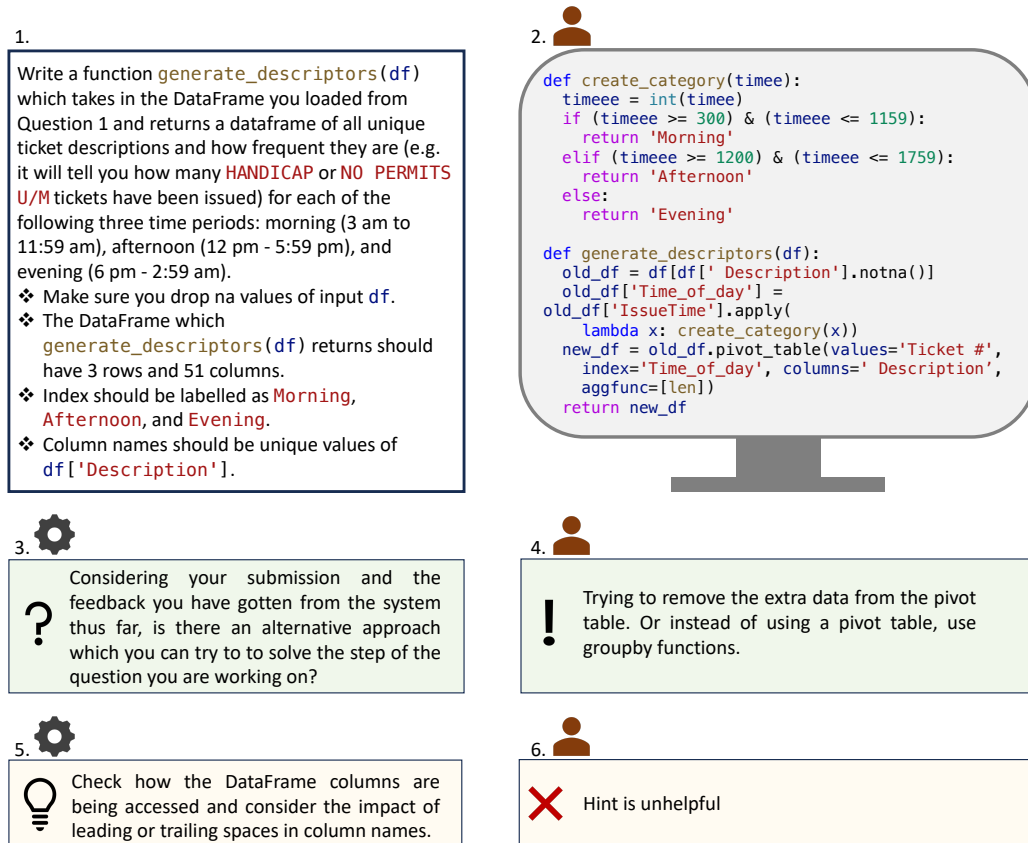


Figure 2: Example illustrating a course question and an interaction between a student in the Before-hint Reflection condition and our support system in Trial 1. The figure shows: 1. Problem description, 2. The student's current buggy program, 3. The system's reflection prompt targeting the evaluation SRL phase, 4. The student's reflection, 5. The AI-generated hint, and 6. The student's evaluation of the hint's helpfulness.

This course is the first mandatory programming course in an online Master's degree program in applied data science at the University of Michigan. Spanning four weeks, the course included weekly assignments, each provided in the form of a JupyterLab notebook containing three to four programming questions, for a total of fourteen questions. Each question required students to complete or modify a given template code to achieve a specific task. Topics covered included regular expressions and dataframe operations using the 'pandas' library. Students could submit assignments multiple times before the weekly deadlines, and their highest assignment scores were counted to-

ward their final grade. Thus, towards the end, almost all students scored 100% on most assignments. The students came from diverse backgrounds but were required to possess basic knowledge of Python programming and statistics. There were 74 students enrolled in the course in the May 2024 semester (Trial 1) and 102 students in the August 2024 semester (Trial 2).

3.2. *AI-generated Hints*

We implemented a technique to automatically generate debugging hints using generative AI, with minor modifications between the two runs of the study. The main idea of the technique was adapted from literature (Phung et al., 2024; Zamfirescu-Pereira et al., 2023). It comprises two main steps. First, the system collected additional relevant information about students’ bugs. This involves (1) obtaining the buggy output by executing the student’s program using the Python interpreter and (2) generating a repaired version of the program by querying a generative AI model. To ensure reliability, the technique verified whether the repaired program correctly solves the task; if it does not, that repaired program was discarded and excluded from the next step. Second, this newly obtained information, together with the student’s buggy program, is inserted into a prompt to query the AI model for a personalized hint. Throughout this process, we used GPT-4 as the AI model.

Similar to the original technique in the literature, the first deployment run (Trial 1) did not incorporate students’ reflections into the hint-generation prompt. However, we observed that some hints in Trial 1 were misaligned with students’ main concerns, as expressed in their reflections, reducing hints’ relevance and quality. Consequently, in the second run (Trial 2), we included students’ reflections (when available) in the prompt to help the AI model focus on the correct issues and produce more targeted, relevant hints. We also made some updates to the language and format of the prompt in the second run. For example, since OpenAI introduced the *system message* feature to better establish context for the query, we modified our prompts to make use of this capability.

3.3. *Reflection Prompts*

In Trial 1, students were randomly assigned to one of three conditions: Before-hint Reflection, After-hint Reflection, or Control (See Figure 1a). Each interaction began when a student clicked the hint-request button for

Table 1: Reflection prompts in Trial 1. For the two experimental conditions, the prompts were randomized per hint request instead of per student: A student could receive different prompts in different requests. There was no prompt for the Control condition.

Condition	Targeted SRL Phase	Reflection Prompt
Before-hint Reflection	Planning	Considering your submission and the feedback you have gotten from the system thus far, what are the steps you think must be followed in order to answer this question, and which step is the one you are currently stuck on?
	Monitoring	Considering your submission and the feedback you have gotten from the system thus far, which topics in the course do you think are most relevant to the current problem you are facing?
	Evaluation	Considering your submission and the feedback you have gotten from the system thus far, is there an alternative approach which you can try to solve the step of the question you are working on?
After-hint Reflection	Planning	Considering the hint you just received and your solution thus far, what steps will you take next to move forward on the question?
	Monitoring	Considering the hint you just received and your solution thus far, are there other topics from the course material you should be incorporating into your solution?
	Evaluation	Considering the hint you just received and your solution thus far, was your general approach a good one, or should you change to an alternative approach to solve the step of the question you are working on?
Control	N/A	N/A

an assignment question. In the Before-hint Reflection condition, students reflected on their progress before receiving a hint (reflection-for-action). In the After-hint Reflection condition, students reflected after receiving and rating a hint (reflection-on-action). The Control group received hints without any reflection prompt.

Three reflection prompts were developed for each experimental condition, with each prompt targeting a SRL phase: planning, monitoring, or evaluation (Zimmerman, 2002). Planning prompts asked students to outline their approach; monitoring prompts encouraged them to connect the task to course material; and evaluation prompts promoted self-assessment and

Table 2: Reflection prompts in Trial 2. Different from Trial 1, in this trial, the prompts were randomized per student: a student always received the same prompt across all hint requests.

Condition	Reflection Prompt
Open	Considering the program you wrote and the feedback you have received from the system so far, what do you think is a possible bug in the program?
Directed	Considering the program you wrote and the feedback you have received from the system so far, what do you think is a possible bug in the program? How does the bug affect the program? What do you think is a way to fix the bug?

consideration of alternatives (See Table 1). These prompts were randomized by hint request, not by student, to examine how prompt type impacted students’ needs during problem-solving. For example, a student could receive a planning prompt for one hint request and an evaluation prompt for another.

In Trial 2, students were randomly assigned to one of the two conditions: Open Prompt or Directed Prompt (See Figure 1b). Depending on their condition, a student was always shown one of the two prompt types (see Table 2). The open prompt asked students to reflect on their progress in general, while the directed prompt guided them to reflect on more specific aspects, such as the impact of the bug on their program and possible solutions.

3.4. Instruments and Data Collection

Our support system consisted of a front-end JupyterLab extension to interact with the students and a back-end server application to execute the hint-generation technique detailed in Section 3.2.

In Trial 1, students needed to click on an Activation button on their JupyterLab and agree to a consent notice to activate our support system. Upon activation, the extension added a "Hint" button below each assignment question. When clicked, it queried the back-end, which generated and returned a hint. Then the system displayed the hint at the top of the screen, providing an option for the student to rate it as either Helpful or Unhelpful. Depending on their condition, students were also prompted to write a reflection, as detailed in Section 3.3 and illustrated in Figure 1a. This reflection practice was optional, and students were allowed to leave the field blank. To prevent excessive use of hints, students were limited to three hint requests per question.

In Trial 2, we made a subtle modification on the front-end: The hint button was shown directly from the beginning, and students could agree to

Table 3: Overview of student participation in Trial 1. Subfigure (a) shows statistics regarding the students: The first row shows the number of students who activated the support system. The second row shows the number of students who have completed at least one interaction with the support system (i.e., receiving a hint for the Control condition or a hint-prompt pair for the experimental conditions). Subfigure (b) shows statistics regarding the hint requests for different conditions and reflection prompts. For space, the SRL phases of planning, monitoring, and evaluation are abbreviated as Plan., Moni., and Eval., respectively. We note that for the After-hint Reflection condition, some hints (8) are not associated with any prompt ($\emptyset$) because in those interactions, students did not rate the hints and thus, were not presented with any prompt. Regarding the second row, an interaction is completed if the student has received both a hint and, for the experimental conditions, a reflection prompt.

(a) Statistics regarding the students

Statistics	Before-hint Reflection	After-hint Reflection	Control	Grand Total
Students activated the support system	12	13	9	34
Students completed at least one interaction	12	13	8	33

(b) Statistics regarding the hint requests

Statistics	Before-hint Reflection				After-hint Reflection					Control	Grand
	Total	Plan.	Moni.	Eval.	Total	Plan.	Moni.	Eval.	$\emptyset$	Total	Total
Hints provided	52	16	17	19	56	17	11	20	8	50	158
Interactions	52	16	17	19	48	17	11	20	0	50	150
Hints rated	50	16	16	18	48	17	11	20	0	41	139

the consent notice and activate hintbot when they request the first hint. This update is aimed at making the system more visible and reducing friction. We still maintain a quota for the number of hints per question, but increase the limit to five hints to accommodate students’ needs. In this trial, all prompts were served before the students received a hint, regardless of the condition. The rest of the interaction workflow remained the same as in Trial 1.

Interacting with the support system was optional, and students were informed that the hints were generated by AI and might not be accurate. We also notified that we collected data on their interactions with the support system, including the timestamps of hint requests, reflection prompts shown, students’ reflections, hints provided, their ratings of hints, along with their assignment submissions. Tables 3 and 4 outline the number of students and their hint requests in our two trials, respectively.

Table 4: Overview of student participation in Trial 2. Similar to Table 3, this table outlines statistics regarding students and hint requests in Trial 2.

Statistics	Open Prompt	Directed Prompt	Grand Total
Students activated the support system	53	45	98
Students completed at least one interaction	37	27	64
Hints provided	236	175	411
Interaction completed	236	175	411
Hints rated	213	168	381

4. Analysis Setup

This section outlines our data analysis procedures for addressing Research Questions 1–3 through two randomized controlled trials. Findings from the Trial 1 were analyzed to answer RQs 1–2, and those from Trial 2 were used to address RQ3.

4.1. RQ1: Impacts of Reflection Prompt and Its Placement

We subdivide RQ1 into four sub-RQs as follows: (a) and (b) examine how reflection practices affect students’ hint-seeking experience and problem-solving performance, respectively. (c) and (d) examine how prompt placements (before-hint versus after-hint) affect students’ participation rates of reflection practices and quality of reflection responses, respectively.

For RQ1a, we measured the hint-seeking experience based on (1) the total number of hint requests made by students throughout the course and (2) their satisfaction with the hints, as indicated by their hint ratings.

For RQ1b, we measured problem-solving performance using two metrics: First, we computed the number of assignment submissions throughout the course. As noted in Section 3.1, by the end of the course, almost all students succeeded in solving all questions. Thus, we used the number of submissions as an indicator of student performance: the fewer the submissions required to achieve their highest scores, the better the performance. Second, we measured performance using the immediate success rate: whether the question was solved in the submission immediately following the hint request.

For RQ1c, we calculated the participation rates for each experimental condition using two metrics: (1) the number of students who wrote at least one non-empty reflection and (2) the number of non-empty reflections.

For RQ1d, we conducted a thematic analysis of students’ written responses to before- and after-hint reflection prompts (Clarke and Braun, 2017; Braun and Clarke, 2012). Thematic analysis is a robust qualitative method for identifying and interpreting themes in textual data, and is well-suited for both inductive and deductive approaches (Clarke and Braun, 2017; Braun and Clarke, 2012; Maguire and Delahunt, 2017). In this study, we adopted a deductive approach focused on describing objective reality (Naeem and colleagues, 2023), theoretically drawing on cyclical SRL model (Zimmerman, 2002) and the 5Rs framework for hierarchical levels of reflection. The 5Rs include Reporting (level 1), Responding, Relating, Reasoning, and Reconstructing (level 5), and served as theoretical grounding to guide the coding and interpretation of student reflections (Bain et al., 1999, 2002; Ryan and Ryan, 2014). According to the 5Rs framework, higher levels of reflection move beyond basic description to the integration of theory and experience, and enables learners to explain, interrogate, and ultimately transform their practice (Ryan and Ryan, 2014; Dhollande et al., 2025). This progression is closely aligned with the development of autonomous learning practices and critical thinking.

The qualitative analysis proceeded in an iterative manner to ensure rigor and reliability. In the first iteration, two authors with strong familiarity with the course independently performed open coding on 10% of the student reflection data, guided by the SRL model (Zimmerman, 2002) and the reflection level framework (Bain et al., 1999, 2002; Ryan and Ryan, 2014). After initial coding, these authors met to discuss their results, collaboratively developed a draft codebook, and identified broader themes. In the second iteration, another author joined the coding team as the third coder. All three coders then independently analyzed the remaining data using the draft codebook, while staying open to new codes as they emerged. For the final step, the three authors worked together to refine the draft codebook (see Table 5). They systematically reviewed the coded data and resolved any discrepancies. This collaborative process ensured both consensus and consistency in the identification of themes across all student reflections. The thematic analysis approach provided a nuanced understanding of how prompt placement influences the depth and quality of student reflection.

4.2. RQ2: Impacts of SRL Phase Associated with Reflection Prompts

We divided RQ2 into four sub-questions to examine the effects of prompts associated with different SRL phases, using the measures established for RQ1,

Table 5: Qualitative codes for thematic analysis of students’ reflections in Trial 1. The creation of these codes is outlined in Section 4.1.

Theme	Code and Description
SRL Phase	Planning (students plan how to approach the problem or set goals for learning)
	Monitoring (students monitor their progress while working on the problem)
	Evaluation (students evaluate their performance)
Critical Engagement	What (Describing key issues/events)
	Why (Explaining why key issues/events are important)
	How (Discussing relevant knowledge they need to resolve issues)
Course Contents	Regular expression (including format/pattern matching, quantifier, lookaheads, list comprehension, <i>VERBOSE</i>)
	<i>pandas</i> (including pivot table, <i>groupby</i> , sort, <i>dropna</i> , <i>merge</i> , <i>concat</i> , <i>filter</i> , array, dataframe, column name change, find the right column)
	<i>NumPy</i> (<i>ndarray</i>)
	<i>Python</i> (<i>for</i> loop)
	Math (percentage)
Hint Assessment	Helpfulness (including the hint is helpful or unhelpful)
	Relevancy (including the hint is relevant or irrelevant)
	Same as the previous hint

including: (a) hint-seeking experience, (b) problem-solving performance, (c) participation rates of reflection practices, and (d) quality of reflection responses.

For this analysis, we omit the After-hint Reflection condition due to a low participation rate of the students in this condition, described more fully in Section 5.1. Due to the small data size, we did not perform statistical tests for RQ2.

Since in Trial 1, prompts were randomly presented for each hint request (see Section 3.3), rather than assigned per student throughout the course, the impacts of prompts could only be measured at the request level, not the student level. Therefore, for the hint-seeking experience in RQ2a, we only measured the hint satisfaction rate, but not the number of hints delivered per student, as in RQ1a. Similarly, for RQ2b about problem-solving per-

formance, we examined the immediate success rates but not the number of submissions per student. For the participation rate in RQ2c, we calculated the number of non-empty reflections. Regarding reflection quality for RQ2d, we analyzed the reflections using the expert-coded annotations described in Section 4.1.

4.3. RQ3: Impacts of Prompt Types in Reflection Practices

RQ3 was addressed through Trial 2. Similar to above, we also subdivided RQ3 into sub-RQs, examining how different types of prompt guidance affect students' (a) hint-seeking experience, (b) problem-solving performance, (c) participation rate of reflection practices, and (d) the quality of reflection responses.

For RQ3a, RQ3b, and RQ3c, we employed the same analytical methods as those applied to RQ1. For RQ3d, we conducted a qualitative coding of student reflection responses using a refined version of the codebook developed during Trial 1. In the initial review of a randomly selected set of 10 responses, the original two coders found it necessary to adapt the Reflection involving SRL phase and Reflection involving Critical Engagement themes, as well as the codes of these themes, to better fit the specific research questions and programming tasks in this study. Because only before-hint reflection prompts were used in RQ3 of Trial 2, the Hint Evaluation theme and its codes, which were originally developed for after-hint reflections in Trial 1, were omitted (see Table 6 for the refined codebook).

After revising the codebook, the same three coders from Trial 1 independently coded a new randomly selected set of 10 responses. Following this round, they discussed and clarified their interpretations of the themes and codes to ensure consistency. Each coder then independently coded 400 responses, after which the coders met to compare their results and reconcile any disagreements. Krippendorff's Alpha (Krippendorff, 2011) was used to assess inter-rater reliability among the three coders for the qualitative coding of RQ3d. For the three Critical Engagement codes (what, why, and how), the reliability values before resolving disagreements were 0.791, 0.649, and 0.670, respectively. After the resolution of discrepancies, these values increased substantially to 0.977, 0.916, and 0.950, indicating excellent agreement among coders.

Table 6: Qualitative codes for the theme "Reflection Quality" used in the analysis of student reflections in Trial 2. Each code is listed with its description. We note that only reflections containing original student text were analyzed. Responses made up solely of copied and pasted error messages were not analyzed. Codes were assigned based on explicit statements in the student reflections, and interpretations were limited to what students actually wrote, without inferring additional meaning.

Theme	Code	Description
Critical Engagement	What	Describing the current error: Student identifies or summarizes the error they are facing (e.g., what is not working, what is incorrect). This is an objective account of an issue encountered at the time of reflection.
	Why	Explaining why the error occurs: Student offers reasons or opinions about the cause of the error or why it made problem-solving challenging. This reflects the student’s own understanding, regardless of whether it is correct or not.
	How	Suggesting how to resolve the error or its cause: Student mentions knowledge, skills, or alternative approaches that could potentially address the error or its root cause. These are strategies the student has not yet attempted at the time of reflection.

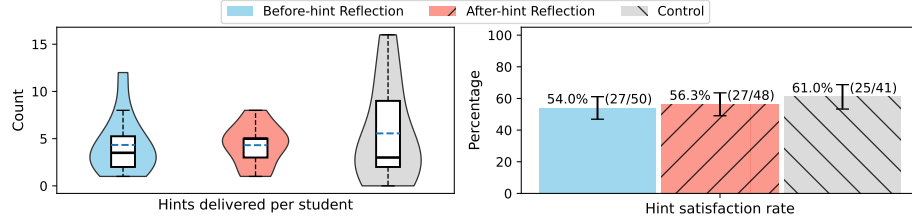
5. Results

This section presents our results following the analysis setup in Section 4.

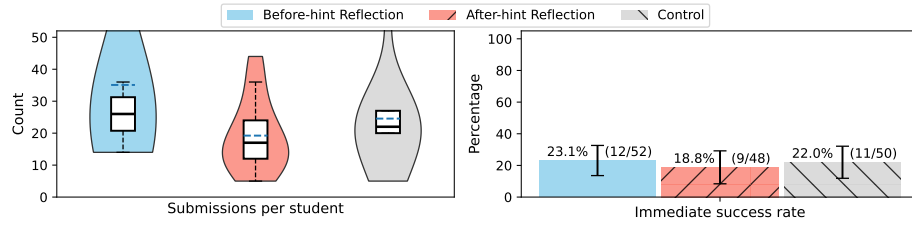
5.1. RQ1: Impacts of Reflection Prompt and Its Placement

5.1.1. RQ1a: Impact of Reflection Practices on Hint-seeking Experience

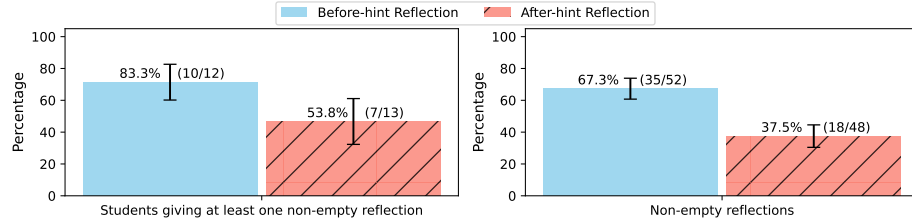
To understand the impact on the hint-seeking experience, we measure the total number of hint requests made by students throughout the course and their satisfaction with the AI-generated hints (see Section 4.1). We calculate the mean number of hint requests per student in each condition. Results indicate a weak relationship. The students in the Control condition requested more hints on average compared to those in the two experimental conditions (see Figure 3a left). However, a two-tailed t-test (Student, 1908) indicates that the difference was not statistically significant ($t = 0.92$, $p = 0.366$, $df = 32$). An ANOVA test (Girden, 1992) to compare the three conditions also results in no significant difference ($F = 0.41$, $p = 0.668$, $df1 = 2$, $df2 = 31$). It is worth noting that the interquartile ranges of the number



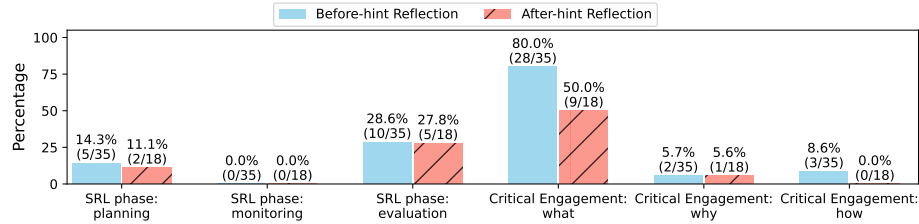
(a) Results for RQ1a: Impacts of reflection practices on students' hint-seeking experience. Hint-seeking experience is measured by the number of hints requested and delivered to the students (left) and their satisfaction rate with the delivered hints (right). In the left plot, the black horizontal lines show the median and the blue dashed lines show the mean.



(b) Results for RQ1b: Impacts of reflection practices on students' solving performance. Solving performance is measured by the total number of submissions made by the students throughout the course (left) and the success rate of solving the questions with the submissions immediately following hint requests (right). In the left plot, the black horizontal lines show median and the blue dashed lines show mean. There is a student in the Before-hint Reflection condition with 146 submissions, and another student in the Control condition with 69 submissions that are not shown due to cutting off of the y-axis for visibility.



(c) Results for RQ1c: Impacts of prompt placement on students' participation rate in reflection practices. The figure shows the percentages of students who wrote at least one non-empty reflection among all students (left) and the percentages of reflections that are non-empty among all reflections (right).



(d) Results for RQ1d: Impacts of prompt placement on students' reflection quality. The figure shows the percentage of non-empty reflections that incorporate different SRL phases and critical engagement, as outlined in Section 4.2 and detailed in Table 5.

Figure 3: Results for RQ1: Impacts of reflection practices and placement.

of requested hints for the experimental conditions are narrower than that of the Control condition.

Next, we examine hint satisfaction rates across conditions by calculating the percentages of hints rated as Helpful rather than Unhelpful. As shown in Figure 3a right, the Before-hint Reflection condition has the lowest mean satisfaction rating at 54.0%, while the Control condition has the highest at 61.0%. However, a χ^2 test suggests no statistically significant differences between these three conditions ($\chi^2 = 0.46$, $p = 0.796$, $df = 2$).

5.1.2. RQ1b: Impact of Reflection Practices on Problem-solving Performance

Problem-solving performance is measured using two metrics: the total number of submissions made and the immediate success rate in solving the questions (see Section 4.1). As shown in Figure 3b left, the Before-hint Reflection condition exhibits the highest number of submissions per student (35.1), followed by the Control condition (24.6) and the After-hint Reflection condition (19.2). An ANOVA test indicates no statistically significant differences across three conditions ($F = 1.36$, $p = 0.271$, $df1 = 2$, $df2 = 31$).

Regarding immediate success, the Before-hint Reflection condition achieves the highest rate (23.1%), closely followed by the Control condition (22.0%), with both being higher than the After-hint Reflection condition (18.8%). A χ^2 test indicates no statistically significant differences across conditions ($\chi^2 = 0.30$, $p = 0.861$, $df = 2$).

5.1.3. RQ1c: Impact of Prompt Placement on Participation Rates

To understand the impact of placement of reflection prompts with regards to hints, we compare participation rates in reflection between the Before-hint and After-hint Reflection conditions by examining the percentage of students who wrote at least one non-empty reflection and the percentage of non-empty reflections overall. As demonstrated in Figure 3c, the Before-hint Reflection condition exhibits higher rates across both metrics. χ^2 tests do not show a significant difference in terms of percentage of students ($\chi^2 = 2.49$, $p = 0.114$, $df = 1$) but indicate significant differences in the percentage of non-empty reflections between the conditions ($\chi^2 = 8.90$, $p = 0.003$, $df = 1$).

5.1.4. RQ1d: Impact of Prompt Placement on Reflection Quality

To understand the impact of reflection prompt placement on reflection practices, we conducted thematic analysis on students' reflection responses to the given prompts (see Section 4.1), found two emerging themes: "SRL

Phases” and “Critical Engagement”. A theme of SRL Phase represents with which specific SRL phases students engaged while responding to each reflection prompt designed to activate one of “planning”, “monitoring”, and “evaluation” phases. Critical Engagement is a theme of critical reflective engagement, which are “What (Describing the current key issues/events)”, “Why (Explaining why the issues are important)”, and “How (Analyzing relevant knowledge they need to resolve the issues).”

Thematic analysis (see Section 4.1) revealed clear differences in the content and focus of student reflections between the two experimental conditions (See Table 7 for examples of reflections and their qualitative code results). Reflections written before receiving hints (before-hint) tended to be more detailed and substantive, frequently referencing specific course concepts, SRL phases, and practical strategies. For instance, one student described planning how to address a data manipulation challenge: "*Trying to remove the extra data from the pivot table. Or [I can] instead of using a pivot table, use groupby functions.*". Others reflected on the challenges of getting started, with comments like "*Right now I need to formulate the problem better I think.*" and "*where to start*" (See Table 7a).

In contrast, most after-hint reflections (11 out of 18 non-empty responses) focused on evaluating the relevance or usefulness of the hints provided, rather than on their own problem-solving process. Typical comments included "*hint was not relevant*" (and "*[The support system] gave me the same hint twice in a row (...)*") (See Table 7b).

Quantitative coding of reflection quality, as shown in Figure 3d, indicated that self-evaluation was the most common SRL phase present in both conditions, while planning and monitoring were less frequently observed. Notably, none of the reflections in either group demonstrated engagement with the monitoring phase. With respect to Critical Engagement codes, the majority of non-empty reflections described the key issues or events students were facing, but fewer than 10% addressed why those issues were important or what knowledge would be needed to resolve them.

5.2. RQ2: Impacts of SRL Phase Associated with Prompts

5.2.1. RQ2a-b: Impact on Hint-seeking Experience and Solving Performance

Hint-seeking experience was measured through hint satisfaction, and problem-solving performance was measured through immediate success rate (see Section 4.2). As shown in Figure 4a, students are most satisfied with hints following the reflective evaluation prompt (66.7%) and least satisfied

Table 7: Examples of reflections written by students in the Before-hint and After-hint Reflection conditions (in Trial 1) with annotation of SRL-Phases and Critical Engagement (Crit. Eng.). The expert annotation process is described in Section 4.1 and the complete list of codes is shown in Table 5. The assignment (A) and question (Q) number followed every response. Coding labels are abbreviated. For SRL phases, P stands for planning, M for monitoring, and E for evaluation. For Critical Engagement, D stands for description of key issues or events (what), E for explanation of why the key issues or events are important (why), and A for analysis of relevant knowledge they need (how).

(a) Before-hint Reflection condition

Before-hint prompt	Reflection	SRL Phase			Crit. Engag.		
		P	M	E	D	E	A
Planning	<i>Right now I need to formulate the problem better I think. (A1Q1)</i>	✓			✓		
Planning	<i>I'm stuck on figuring out the regex that distinguishes the text in the user-name column. (A1Q3)</i>			✓	✓	✓	
Monitoring	<i>Consolidating excel files into one and merging dataframes (A3Q1)</i>			✓	✓		
Monitoring	<i>loading data says load_ticket_data is not a dataframe when it is pd.concat (A3Q1)</i>				✓		
Evaluation	<i>Perhaps I should create a for loop that adds each match to a list? (A1Q1)</i>				✓		✓
Evaluation	<i>Trying to remove the extra data from the pivot table. Or instead of using a pivot table, use groupby functions (A3Q2)</i>	✓		✓	✓		

(b) After-hint Reflection condition

After-hint prompt	Reflection	SRL Phase			Crit. Engag.		
		P	M	E	D	E	A
Planning	<i>I should be trying to use the VERBOSE form, but I cannot get it to work in that format (A1Q3)</i>				✓		
Planning	<i>hint was not relevant (A3Q1)</i>						
Monitoring	<i>I think my approach is solid; in fact manually checking my results yields the same answer as my function. I'll parse through the data more; maybe there are letter case issues I haven't considered (A3Q4)</i>	✓		✓	✓		
Monitoring	<i>This hint was both helpful and non-helpful – helpful as in it's correct, but unhelpful in that I already realized that and that is what I'm spinning my wheels trying to do (A4Q4)</i>						
Evaluation	<i>helpful (A4Q3)</i>						
Evaluation	<i>I don't think IssueTime is the issue I'm having, I'm struggling with the pivot table (A3Q2)</i>			✓	✓		

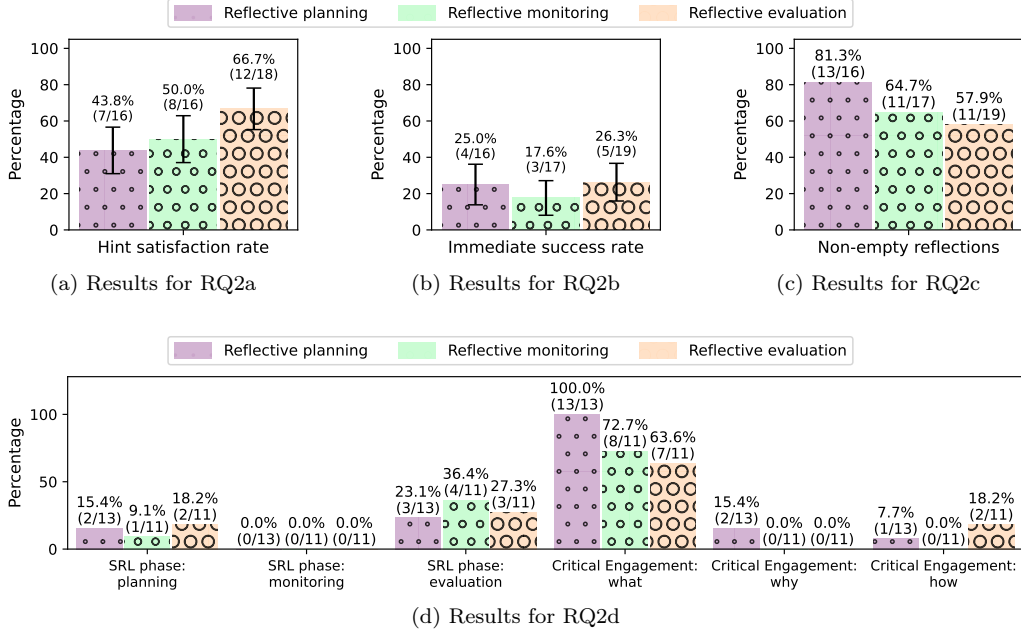


Figure 4: Results for RQ2: Impacts of SRL phase associated with reflection prompts. Similar to Figure 3, this figure shows the impacts of the SRL phase associated with reflection prompts on: (a) hint satisfaction rate, (b) immediate success rate, (c) non-empty reflection rate, and (d) reflection quality.

with those following the reflective planning prompt (43.8%). The immediate success rate, on the other hand, is highest for the evaluation prompt (26.3%) and lowest for the monitoring prompt (17.6%), as shown in Figure 4b.

5.2.2. RQ2c-d: Impact on Participation Rate and Reflection Quality

The results for RQ2c and RQ2d are shown in Figures 4c and 4d, respectively. We observe varying participation rates, with students writing the most non-empty reflections for the reflective planning prompt (81.3%) and the fewest for the reflective evaluation prompt (57.9%). This pattern also holds in the qualitative coding, in the Critical Engagement code related to the description of key issues or events. Notably, all 13 non-empty reflections following the reflective planning prompt contain a description of the key issues or events the students were facing (examples are shown in Figure 7a).

5.3. RQ3: Impacts of Prompt Types on Reflection Practices

5.3.1. RQ3a-b: Impact on Hint-seeking Experience and Solving Performance

As shown in Figure 5a, the number of hints per student does not differ much across the two prompt conditions, with the Open Prompt condition slightly higher with 4.5 hints per student compared to the Directed Prompt condition with 3.9. A two-tailed t-test indicates no difference between the two conditions ($t = 0.45$, $p = 0.650$, $df = 96$). In terms of hint satisfaction rate, students in the Open Prompt condition rated hints more highly. A χ^2 test suggests a statistically significant difference in hint rating ($\chi^2 = 5.71$, $p = 0.017$, $df = 1$).

Figure 5b illustrates the results for RQ3b. A two-tailed t-test on the number of submissions ($t = -1.02$, $p = 0.312$, $df = 96$) and a χ^2 test on immediate success rate ($\chi^2 = 0.09$, $p = 0.769$, $df = 1$) suggest no difference between the two conditions.

5.3.2. RQ3c-d: Impact on Participation Rate and Reflection Quality

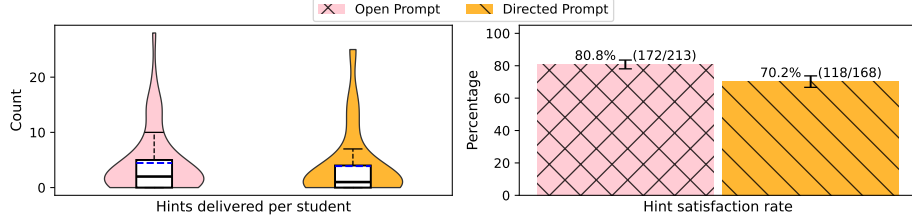
Figure 5c shows the results for RQ3c. Two χ^2 tests on the number of students giving non-empty reflections ($\chi^2 = 0.07$, $p = 0.794$, $df = 1$) and the number of non-empty reflections ($\chi^2 = 1.22$, $p = 0.270$, $df = 1$) suggest no difference between the two conditions.

For RQ3d, Figure 5d summarizes the qualitative coding results for student reflection as responses to the directed and open prompts. Each reflection was evaluated for the presence of three components: “what” (describes the current error), “why” (explains why the error occurs), and “how” (suggests how to resolve the error or its cause). Examples of student reflection to the Open and Directed Prompt types are presented in Tables 8a and 8b, respectively.

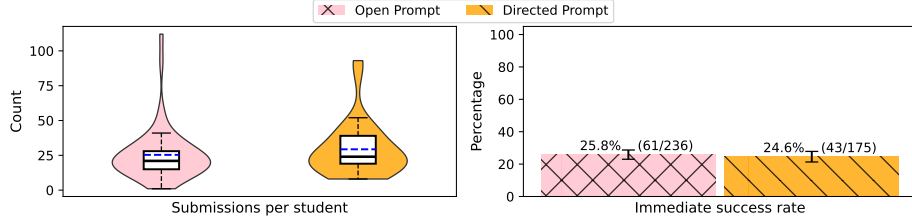
To examine differences in the frequency of each qualitative code between the prompt types, χ^2 tests were conducted. For the “what” code ($\chi^2 = 2.20$, $p = 0.138$, $df = 1$) and the “why” code ($\chi^2 = 0.00$, $p = 1.000$, $df = 1$), no significant difference was observed. However, for the “how” code, a significant difference was found between conditions ($\chi^2 = 9.73$, $p = 0.002$, $df = 1$).

6. Discussion

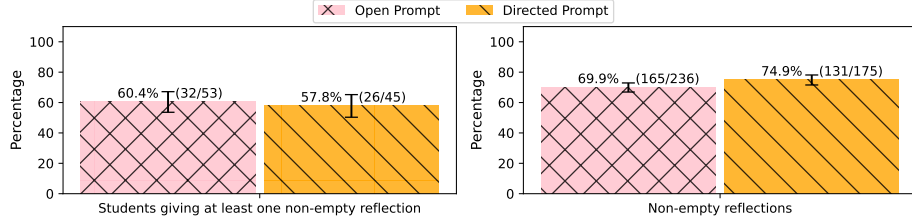
In this section, we discuss the insights and implications for AI design and development in educational settings. First, we will go through the primary insights and implications based on our overall findings, and then zoom down on insights and implications from answers to specific RQs.



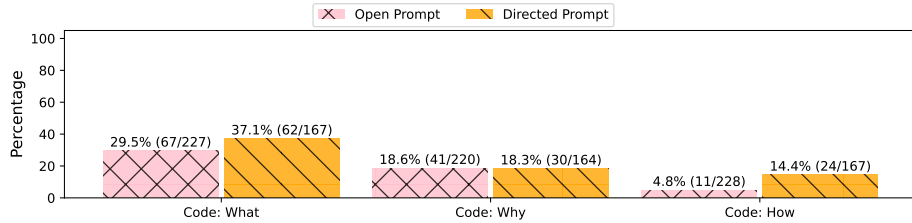
(a) Results for RQ3a: Similar to Figure 3a, this figure shows the impact of prompt type on students' help-seeking experience.



(b) Results for RQ3b: Similar to Figure 3b, this figure shows the impact of prompt type on students' solving performance.



(c) Results for RQ3c: Similar to Figure 3c, this figure shows the impact of prompt type on students' participation in reflection practices.



(d) Results for RQ3d: Similar to Figure 3d, this figure shows the impacts of prompt type on students' reflection quality. The codes are outlined in Section 4.3 and detailed in Table 6. We note that unresolved cases where the coders could not reach consensus are not included in the analysis: $N_{what} = 6$ (six cases where coders could not agree whether "what" should be applied.), $N_{why} = 16$, $N_{how} = 5$.

Figure 5: Results for RQ3: Impacts of prompt types on students' hint-seeking experience and solving performance.

Table 8: Examples of reflections written by students in the Open and Directed Prompt conditions (in Trial 2) with annotation of theme Critical Engagement. The expert annotation process is described in Section 4.3 and the complete list of codes is shown in Table 6. The assignment (A) and question (Q) number followed every response.

(a) Open Prompt condition			
Reflection	Code		
	What	Why	How
<i>[Student' regex] pattern needs work? (A1Q3)</i>			
<i>My dictionary might not be saved correctly, as I'm getting an assert error that I'm not returning a dict (A2Q1)</i>	✓	✓	
<i>I still cannot find the column names you're looking for. I just don't know what they are. What are they? I won't be able to find out otherwise (A2Q3)</i>			
<i>could you take a look at the codes for question 1? I am having a hard time to match out the total number of rows (A3Q1).</i>	✓		
<i>What could be wrong with my percentage delays? United airlines isn't the bottom one (A4Q3)</i>	✓	✓	
<i>What's wrong? (A4Q3)</i>			
(b) Directed Prompt condition			
Reflection	Code		
	What	Why	How
<i>I received the following error: You have failed this test due to an error. The traceback has been removed because it may contain hidden tests. This is the exception that was thrown: AssertionError: Sorry, this item should be in the log results, check your formatting. Does this mean I have stored something in the wrong variable? (A1Q3)</i>		✓	
<i>is my answer correct just with a negative instead of positive correlation? i'm pretty sure my code is right! (A2Q3)</i>		✓	
<i>I believe my code is correct, It has the right shape and when i inspect it through snapshots it seems to be right. does it look ok (A3Q1)</i>			✓
<i>I have all the correct counts except vanity_count. I believe that the error comes in when I calculate total_plates. I don't think I should have the .notna() function, but the code doesn't run without it (A3Q4)</i>	✓	✓	✓
<i>i am lost, i am getting an error without a specific reason (A4Q3)</i>			
<i>One bug that I can easily identify is the shape, but I can't tell which row is extraneous, considering I'm expecting 11,5 and getting 12,5. My biggest question is why I'm only getting NaN values in my resulting dataframe. (A4Q4)</i>	✓	✓	

6.1. Tension between Pedagogical Values and AI Optimization Methods

6.1.1. Summary of the Findings

Across all three research questions, our findings reveal a tension between student satisfaction with AI-generated hints and the quality of their reflection. Conditions that yielded higher satisfaction consistently corresponded to lower-quality reflection practices, as measured through hint satisfaction rates, reflection depth, and participation metrics. In RQ1, students who were prompted to reflect after receiving hints (after-hint) demonstrated higher (though not statistically significant) satisfaction compared to the before-hint group. However, the before-hint group exhibited significantly more higher participation rate of reflection and demonstrated more advanced reflection levels. For RQ2, the reflective evaluation prompt resulted in the highest hint satisfaction but the lowest rates of both meaningful and advanced reflections. Conversely, the reflective planning prompt led to the lowest satisfaction but the highest rates of participation and responses with more sophisticated reflection components. In RQ3, students in the open prompt condition reported significantly higher hint satisfaction rates compared to the directed prompt condition. While differences in participation rate were not statistically significant, students in the directed prompt type were statistically significantly more likely to produce constructive reflection responses that included deeper reflection components (“how”), despite reporting significantly lower satisfaction with the hints.

On the other hand, no clear patterns were found between the immediate success rate and either hint satisfaction or reflection quality.

6.1.2. Theoretical Interpretation

The inverse relationship between quality reflection and satisfaction with AI-generated hints echoes established research on the dynamics between effortful learning and satisfaction of students. Reflection, as a component of SRL, demands substantial cognitive resources (de Bruin et al., 2020; Seufert et al., 2024). Previous research demonstrates that student satisfaction generally decreases as mental effort increases, especially when cognitive load exceeds manageable levels (Bradford, 2011; Chen et al., 2023). Bradford (2011) found that cognitive load accounted for about 25% of the variance in satisfaction in online learning. Similarly, Chow and Yeh (2022) found that students showed diminished satisfaction and reduced willingness to invest cognitive effort in tasks they perceived as too easy or too difficult. A de-

crease in satisfaction and willingness on learning tasks was likely to develop into disengagement with learning activities.

Existing SRL theoretical frameworks focusing on affective aspects could provide important foundations to balance effortful learning with AI assistance for students to continue engaging meaningfully with learning activities. Efklides’ Metacognitive and Affective model of SRL (MASRL) (Efklides, 2011) emphasizes the interactions between metacognition, motivation, and affect across two levels of functioning: the Person level (stable characteristics) and the Task X Person level (dynamic task-specific interactions). Boekaerts’ Dual Processing model (Boekaerts, 2011) describes how students navigate between growth pathways and well-being pathways based on their appraisals of learning situations, including their emotional responses. Integrating those SRL frameworks into educational research in technology-rich educational contexts would add value by advancing structured understanding of ways to embed AI assistance for more benefits and less harm on student SRL.

6.1.3. Systemic Implication

Our findings, combined with educational theories, call for systemic changes in how AI models are trained, evaluated, and aligned for educational use. Current AI development practices primarily optimize for user satisfaction (Ouyang et al., 2022; Bai et al., 2022). However, user preferences in education often favor convenience and ease over pedagogical value. As a result, AI systems may perform tasks that allow learners to bypass cognitive effort, even when effort is essential for skill development.

This finding highlights the importance of educational AI system designs that incorporate the concept of desirable difficulties rather than solely pursuing user satisfaction maximization (Bjork, 1994; Bjork et al., 2011). Prioritizing satisfaction alone in AI optimization risks diminishing opportunities for students to ‘struggle’ productively and develop SRL skills (e.g., reflection) and other fundamental learning competencies such as critical thinking. Many studies have already shown that most learning occurs as students are challenged; they learn as they productively struggle (Warshauer, 2015a,b), work on tasks with desirable difficulty (Bjork et al., 2011) that students independently cannot resolve but can with others’ assistance (Vygotsky and Cole, 1978). It is a timely issue to move to a focus on inferring and providing the zone of proximal difficulty as industry leaders, researchers, and practitioners examine, adopt, and create AI-powered educational tools. While the majority of AI-powered tools are developed to effectively assist students, our study

findings draw our attention to AI development to sometimes have students struggle.

Our study findings also call for the transformation of the learning evaluation frameworks. Our results show negligible effects on immediate success rates, a metric used for problem-solving performance. Under the current practices of summative evaluation of short-term performance, the trade-off between learning opportunity and satisfaction would not be easily detected. Thus, it is important for AI researchers in educational settings to adopt diverse metrics beyond performance to evaluate the impact of AI use on learning (Bjork, 1994; Bjork et al., 2011; Coulson and Harvey, 2012).

In our study, we mitigated this problem through context engineering: explicitly signaling the educational intent in the AI-prompt and designing non-conversational user interactions to limit pedagogically irrelevant information (such as prompt injection attacks (Greshake et al., 2023; Perez and Ribeiro, 2022)). While effective to some extent, such design choices also limit interaction depth. A more scalable solution could involve pre-training foundational models on pedagogically aligned data. This would allow for more robust multi-turn dialogue with learners while maintaining a focus on learning goals, not just user satisfaction (Christiano et al., 2017; Stiennon et al., 2020).

Foundation models are typically trained on large, diverse text datasets and aligned with human preferences using techniques such as Direct Preference Optimization (DPO) (Rafailov et al., 2023) and Reinforcement Learning through Human Feedback (RLHF) (Ouyang et al., 2022). Multiple methods exist for preference optimization, including ordinal ranking of responses (Christiano et al., 2017), binary feedback through user voting mechanisms, as employed in this study and typical in conversational systems with thumbs up/down rating features (Stiennon et al., 2020), and AI-based response scoring for safety guardrails (Bai et al., 2022). However, none of these methods address the subjective qualities underlying preference definitions. Current preference alignment approaches generally rely on end-user opinions, and while correctness is often used as a preference criterion (Stiennon et al., 2020), theories of metacognition (Efklides, 2011), SRL (Zimmerman, 2011), and productive struggle (Bjork, 1994) are not integrated into model alignment processes. Integrating these theoretical constructs into the model alignment process could lead to AI systems better suited for long-term learning outcomes rather than short-term satisfaction.

In short, our findings reveal a fundamental tension between AI optimiza-

tion objectives and pedagogical effectiveness. It demands reconsideration of how AI systems are designed and trained for educational contexts. The evidence suggests that prioritizing immediate user satisfaction may undermine the deeper learning processes that educational AI systems should support.

6.1.4. *Limitations and Future Work*

Several limitations must be noted. Our study was conducted in a fully on-line programming course with adult learners. Future research should examine how these dynamics play out with K-12 students, who are still developing SRL and higher-order thinking skills. We disabled conversational interaction between students and the AI to avoid gaming behaviors. This design choice limited our ability to explore how dialogue types might affect student satisfaction-reflection dynamics, such as the impacts of AI response types on students' metacognition and cognition activities. Our study did not incorporate human-in-the-loop designs. Exploring how human instructors and tutors might ease the discovered tension and complement our AI-powered tool remains a critical area for future work.

6.2. *The Impact of Prompt Types with AI-Generated Hints*

6.2.1. *Summary of the Findings*

In Trial 1, addressing RQ1a-b, students in both experimental groups (prompts *before* and *after* hints) tended to request fewer AI-generated hints than the Control group (no prompt), although this difference was not statistically significant. However, the experimental groups exhibited narrower interquartile ranges in the number of hint requests, suggesting more consistent behavior. No significant differences were found in immediate success rates or assignment submission counts across conditions. Between *before*- and *after*-hint conditions (RQ1c-d), prompting students to reflect *before* receiving AI-generated hints led to higher participation with reflection practices, more constructive reflection responses, and lower satisfaction with the hints. In contrast, students prompted *after* receiving hints often skipped the prompt or gave brief, superficial comments (e.g., “helpful”).

For RQ2, although no statistical tests were performed due to a small data size, trends showed that for the *before*-hint condition, the reflective *evaluation* prompt was associated with the highest satisfaction with hints. However, it led to the lowest participation rate in reflection, while the *planning* prompt resulted in the highest participation.

Regarding RQ3, *open* prompt group students produced less constructive reflection responses compared to *directed* prompts; their responses often had a description of “what” the current error was, but fewer provided analysis of “why” the error occurred or “how” to address it.

6.2.2. Theoretical Interpretation

In our study, experimental groups compared to Control (RQ1) and directed prompt groups compared to open prompts (RQ3) tended to ask for a lower number of hints. Yet, the difference was not significant. Previous literature reported mixed effects of reflection on help-seeking frequency, which were often null effects or a decrease in help-seeking frequency (Seufert et al., 2024; Janssen et al., 2023; Mandernach et al., 2011; Broadbent and Poon, 2015). The narrower interquartile ranges of both before- and after-hint groups might be worthy of further investigation. One hypothesis to explain the results would be that confounding variables, such as relevant skill competencies or prior knowledge, could have interfered with the impact of the conditions on the number of hint requests. The impact of the conditions might have still decreased the need for more hints of struggling students, resulting the narrower ranges compared to the Control group.

Impacts on reflection quality display clearer patterns throughout trials. *After*-hint reflection might have been not as effective as *before*-hint reflection (RQ1). This is because exposure to hints prior to reflection might have led students to focus more on the hint content rather than pause to analyze their current problems in their own words (Baars et al., 2020; Janssen et al., 2023; Coulson and Harvey, 2012; Moon, 2004). In addition, some students might have thought that their reflection responses would be used as input to generate more personalized AI-generated hints, even though our instructions clearly conveyed that their response was not used for hint generation. While there is not much research on direct comparison between planning and evaluation in the context of reflection prompting, one potential explanation would be that *planning* might have felt more actionable and aligned with a type of assistance for which students sought. While *planning* and *evaluation* are not two clearly separate phases, *planning* prompts might have been more associated with progress and pathway forward (e.g., “(...) what are the steps you think must be followed in order to answer this question (...),” similar with what they expected from hints. This pattern is consistent with previous research, which has shown that students are more likely to engage in lower-level reflections, such as description, than in higher-level reflections

including integration and interpretation (Bain et al., 1999, 2002; Ryan and Ryan, 2014). On the other hand, it might have felt more challenging for students to *evaluate* their problems and approaches at which they were stuck (“(...) is there an alternative approach which you can try to solve the step of the question (...)”). Additionally, students are shown to find it easier to reflect on current challenges than to evaluate alternatives, which can be more cognitively demanding (Wirth, 2009; Lin, 2022). Finally, the directed prompt with more specific guidance (RQ3) could have been more effective at leading students toward quality reflection, especially to students who were not familiar with reflections and more likely to be tempted to metacognitive laziness to avoid frustration (Wirth, 2009; Lin, 2022). Previous research showed that open prompts are less beneficial for students with lower autonomous learning capabilities (Wirth, 2009; Hadwin et al., 2011).

6.2.3. *Practical Implications*

These findings underscore the importance of well-designed reflection interventions, especially when students are experiencing high cognitive load and AI-powered tools are a click away. Well-designed reflection prompting could guide students to deliberately pause to engage with reflections, an effortful and beneficial learning process, instead of hasty application of hints. Furthermore, students could also practice and integrate higher-level reflection (analysis of “why” they are stuck or “how” to integrate hints to resolve their problems) to their real-world learning processes (Bain et al., 1999, 2002; Ryan and Ryan, 2014). Specifically, this study informs the decision of when to provide hints, which SRL phases to target, and how much guidance to provide.

6.2.4. *Limitations and Future Work*

This study did not measure the long-term development of students’ reflection skills or performance. Future research should adopt a longitudinal design to assess how reflection practices combined with AI-generated hints influence reflective growth and learning outcomes. Due to our sample size and the measures collected, definitive conclusions about when and how reflection practices affect AI support usage and learning gains remain elusive. Future studies with larger, more diverse cohorts should investigate whether reflection prompts can consistently reduce students’ dependence on AI assistance, elevate satisfaction through metacognitively informed hinting, and foster equitable SRL across ability levels. Finally, interactions between reflec-

tion prompting and student competency warrant closer examination. This understanding could clarify under which conditions and for which students reflection practices meaningfully impact help-seeking behavior.

7. Conclusions

This study advances understanding of how reflective scaffolding can be integrated with AI-generated hints to shape student learning in programming courses. Our findings reveal that while reflection before receiving hints and the use of planning-focused or directed prompts foster higher-quality reflection, these approaches are often associated with lower student satisfaction. Moreover, a consistent negative relationship emerged between satisfaction and reflection quality, and immediate problem-solving performance was largely unaffected by reflection interventions. These results highlight a fundamental tension between optimizing for user satisfaction and promoting deeper, self-regulated learning. Moving forward, educational AI systems should be intentionally designed to prioritize pedagogical value and support the development of students' critical thinking and independent learning skills, rather than focusing solely on satisfaction metrics. This work provides insights for researchers, educators, instructional designers, and AI developers seeking to align AI-powered support with long-term educational goals.

References

- Aleven, V., McLaren, B., Roll, I., Koedinger, K., 2006. Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education* 16, 101–128.
- Baars, M., Wijnia, L., de Bruin, A., Paas, F., 2020. Training self-regulation: The effects of self-regulation training on learning and performance. *Learning and Instruction* 68, 101325.
- Bai, Y., Kadavath, S., Kundu, S., Askill, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al., 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* .
- Bain, J.D., Ballantyne, R., Packer, J., Mills, C., 1999. Using journal writing to enhance student teachers' reflectivity during field experience placements. *Teach. Teach.* 5, 51–73.

- Bain, J.D., Mills, C., Ballantyne, R., Packer, J., 2002. Developing reflection on practice through journal writing: Impacts of variations in the focus and level of feedback. *Teach. Teach.* 8, 171–196.
- Bardach, L., Klassen, R.M., Durksen, T.L., Rushby, J.V., Bostwick, K.C., Sheridan, L., 2021. The Power of Feedback and Reflection: Testing an Online Scenario-based Learning Intervention for Sstudent Teachers. *Computers & Education* 169.
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, O., Mariman, R., 2024. Generative AI Can Harm Learning. Available at SSRN 4895486.
- Becker, B.A., Denny, P., Finnie-Ansley, J., Luxton-Reilly, A., Prather, J., Santos, E.A., 2023. Programming Is Hard - Or at Least It Used to Be: Educational Opportunities and Challenges of AI Code Generation, in: *Proceedings of the Technical Symposium on Computer Science Education (SIGCSE)*, pp. 500–506.
- Bjork, E.L., Bjork, R.A., et al., 2011. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society* 2, 56–64.
- Bjork, R.A., 1994. Memory and metamemory considerations in the training of human beings. *Metacognition: Knowing about Knowing* , 185–205.
- Boekaerts, M., 2011. Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology* 54, 199–231.
- Boud, D., 2013. *Enhancing Learning through Self-Assessment*. Routledge.
- Boud, D., Keogh, R., Walker, D., 1985. What is reflection in learning. *Reflection: Turning experience into learning* , 7–17.
- Bradford, G.R., 2011. A relationship study of student satisfaction with learning online and cognitive load: Initial results. *The Internet and Higher Education* 14, 217–226.
- Braun, V., Clarke, V., 2012. *Thematic Analysis*. American Psychological Association.

- Broadbent, J., Poon, W.Y., 2015. Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *Internet and Higher Education* 27, 1–13.
- de Bruin, A.B., Baars, M., van Merriënboer, J.J., 2020. The effort monitoring and regulation (emr) model: A systematic review of the literature. *Educational Psychology Review* 32, 903–926.
- Burner, T., Lindvig, Y., Wærness, J.I., 2025. “We Should Not Be Like a Dinosaur”—Using AI Technologies to Provide Formative Feedback to Students. *Education Sciences* 15.
- Chen, Y., Wang, Q., Chen, J., 2023. Investigating students’ satisfaction with online collaborative learning: The role of cognitive load. *Computers & Education* 201, 104843.
- Choi, H., Jovanovic, J., Poquet, O., Brooks, C., Joksimovic, S., Williams, J.J., 2023. The Benefit of Reflection Prompts for Encouraging Learning with Hints in an Online Programming Course. *The Internet and Higher Education* 58.
- Chow, N.C.H., Yeh, I.J., 2022. Correlation between learning motivation and satisfaction in synchronous on-the-job online training in the public sector. *Frontiers in Psychology* 13, 789252.
- Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D., 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems* 30, 4299–4307.
- Clarke, V., Braun, V., 2017. Thematic analysis. *The Journal of Positive Psychology* 12, 297–298.
- Coulson, D., Harvey, M., 2012. Scaffolding student reflection for experience-based learning: A framework. *Teaching in Higher Education* 18, 401–413.
- Davis, E.A., 2003. Prompting middle school science students for productive reflection: Generic and directed prompts. *The Journal of the Learning Sciences* 12, 91–142.
- Davis, E.A., Linn, M.C., 2000. Scaffolding students’ knowledge integration: Prompts for reflection in kie. *International Journal of Science Education* 22, 819–837.

- Denny, P., Gulwani, S., Heffernan, N.T., Käser, T., Moore, S., Rafferty, A.N., Singla, A., 2024a. Generative AI for Education (GAIED): Advances, Opportunities, and Challenges. CoRR abs/2402.01580.
- Denny, P., MacNeil, S., Savelka, J., Porter, L., Luxton-Reilly, A., 2024b. Desirable Characteristics for AI Teaching Assistants in Programming Education, in: Proceedings of the Innovation and Technology in Computer Science Education (ITiCSE), pp. 408–414.
- Dewey, J., 1997. How We Think.
- Dhollande, P., Kim, J., Lee, S., 2025. Fostering autonomous learning practices through reflective frameworks. *Journal of Learning Analytics* 12, 50–68.
- Edwards, S., 2017. Reflecting differently. new dimensions: reflection-before-action and reflection-beyond-action. *International Practice Development Journal* 7.
- Edwards, S.H., 2004a. Using software testing to move students from trial-and-error to reflection-in-action, in: Proceedings of the 35th SIGCSE technical symposium on Computer science education, pp. 26–30.
- Edwards, S.H., 2004b. Using software testing to move students from trial-and-error to reflection-in-action, in: Proceedings of the 35th SIGCSE technical symposium on Computer science education, ACM, New York, NY, USA.
- Efklides, A., 2011. Interactions of metacognition with motivation and affect in self-regulated learning: The masrl model. *Educational Psychologist* 46, 6–25.
- Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., Gašević, D., 2025. Beware of Metacognitive Laziness: Effects of Generative Artificial Intelligence on Learning Motivation, Processes, and Performance. *British Journal of Educational Technology* 56.
- Finnie-Ansley, J., Denny, P., Becker, B.A., Luxton-Reilly, A., Prather, J., 2022. The Robots Are Coming: Exploring the Implications of OpenAI Codex on Introductory Programming, in: Australasian Computing Education Conference (ACE), pp. 10–19.

- Gerlich, M., 2025. Ai tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies* 15, 6.
- Girden, E.R., 1992. ANOVA: Repeated Measures.
- Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., Fritz, M., 2023. Not what you’ve signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. *arXiv preprint arXiv:2302.12173* .
- Hadwin, A.F., Järvelä, S., Miller, M., 2011. Self-regulated learning in the digital age: A theoretical review. *Educational Psychologist* 46, 267–284.
- Handa, K., Bent, D., Tamkin, A., McCain, M., Durmus, E., Stern, M., Schiraldi, M., Huang, S., Ritchie, S., Syverud, S., Jagadish, K., Vo, M., Bell, M., Ganguli, D., 2025. Anthropic education report: How university students use claude.
- Janssen, E.M., Lossius, M.H., Muis, K.R., 2023. Overcoming misconceptions about desirable difficulties in learning. *Educational Psychology Review* 35, 1–28.
- Johnson, C.I., Mayer, R.E., 2010. Applying the self-explanation principle to multimedia learning in a computer-based game-like environment. *Computers in Human Behavior* 26, 1246–1252. doi:10.1016/j.chb.2010.03.025.
- Kosmyna, N., Hauptmann, E., Yuan, Y.T., Situ, J., Liao, X.H., Beresnitzky, A.V., Braunstein, I., Maes, P., 2025. Your Brain on Chatgpt: Accumulation of Cognitive Debt When Using an Ai Assistant for Essay Writing Task. *ArXiv Preprint ArXiv:2506.08872* .
- Krippendorff, K., 2011. Computing krippendorff’s alpha-reliability .
- Kumar, H., Xiao, R., Lawson, B., Musabirov, I., Shi, J., Wang, X., Luo, H., Williams, J.J., Rafferty, A.N., Stamper, J.C., Liut, M., 2024. Supporting Self-Reflection at Scale with Large Language Models: Insights from Randomized Field Experiments in Classrooms, in: *Proceedings of the Conference on Learning @ Scale, (L@S)*, pp. 86–97.

- Lee, H.P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., Wilson, N., 2025. The impact of generative ai on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers, in: Proceedings of the 2025 CHI conference on human factors in computing systems, pp. 1–22.
- Lin, G., 2022. Using metacognitive prompts to enhance self-regulated learning and learning outcomes: A meta-analysis of experimental studies in computer-based learning environments. *Journal of Computer Assisted Learning* 38, 811–832.
- Lishinski, A., Yadav, A., 2021a. Self-evaluation interventions: Impact on self-efficacy and performance in introductory programming. *ACM Transactions on Computing Education (TOCE)* 21, 1–28.
- Lishinski, A., Yadav, A., 2021b. Self-evaluation interventions: Impact on self-efficacy and performance in introductory programming: Impact on self-efficacy and performance in introductory programming. *ACM Trans. Comput. Educ.* 21, 1–28.
- Loksa, D., Xie, B., Kwik, H., Ko, A.J., 2020. Investigating novices’ in situ reflections on their programming process, in: Proceedings of the 51st ACM technical symposium on computer science education, pp. 149–155.
- Ma, B., Chen, L., Konomi, S., 2024. Enhancing Programming Education with Chatgpt: A Case Study on Student Perceptions and Interactions in a Python Course, in: Proceedings of the Artificial Intelligence in Education (AIED), pp. 113–126.
- Maguire, M., Delahunt, B., 2017. Doing a Thematic Analysis: A Practical, Step-by-Step Guide for Learning and Teaching Scholars. *All Ireland Journal of Higher Education* 9.
- Mandernach, B.J., Forrest, K., Babutzke, J., Manker, L., 2011. The effects of student engagement, student satisfaction, and perceived learning in online learning environments. *Online Learning* 15, 74–88.
- Mann, K., Gordon, J., MacLeod, A., 2009. Reflection and Reflective Practice in Health Professions Education: a Systematic Review. *Advances in Health Sciences Education* 14.

- Margulieux, L.E., Prather, J., Reeves, B.N., Becker, B.A., Uzun, G.C., Loksa, D., Leinonen, J., Denny, P., 2024. Self-Regulation, Self-Efficacy, and Fear of Failure Interactions with How Novices Use LLMs to Solve Programming Problems, in: Proceedings of the Innovation and Technology in Computer Science Education (ITiCSE), pp. 276–282.
- Marwan, S., Williams, J.J., Price, T.W., 2019. An Evaluation of the Impact of Automated Programming Hints on Performance and Learning, in: Conference on International Computing Education Research (ICER), pp. 61–70.
- Moon, J.A., 2004. A Handbook of Reflective and Experiential Learning: Theory and Practice. Routledge.
- Muñoz-Merino, P.J., Valiente, J.A.R., Kloos, C.D., 2013. Inferring higher level learning information from low level data for the khan academy platform, in: Proceedings of the third international conference on learning analytics and knowledge, pp. 112–116.
- Naeem, M., colleagues, 2023. A pragmatic approach to thematic analysis. *International Journal of Qualitative Methods* 22, 1–12.
- OpenAI, 2023. ChatGPT. <https://openai.com/blog/chatgpt>.
- O’Rourke, R., 1998. The learning journal: from chaos to coherence. *Assessment & Evaluation in Higher Education* 23, 403–413.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al., 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35, 27730–27744.
- Pankiewicz, M., Baker, R.S., 2023. Large Language Models (GPT) for Automating Feedback on Programming Assignments. *CoRR* abs/2307.00150.
- Pankiewicz, M., Baker, R.S., 2024. Navigating Compiler Errors with AI Assistance - A Study of GPT Hints in an Introductory Programming Course, in: Proceedings of Innovation and Technology in Computer Science Education (ITiCSE), pp. 94–100.

- Perez, F., Ribeiro, I., 2022. Ignore previous prompt: Attack techniques for language models. arXiv preprint arXiv:2211.09527 .
- Phung, T., Pădurean, V., Cambronero, J., Gulwani, S., Kohn, T., Majumdar, R., Singla, A., Soares, G., 2023. Generative AI for Programming Education: Benchmarking Chatgpt, GPT-4, and Human Tutors, in: Proceedings of the Conference on International Computing Education Research (ICER) - Volume 2, pp. 41–42.
- Phung, T., Padurean, V., Singh, A., Brooks, C., Cambronero, J., Gulwani, S., Singla, A., Soares, G., 2024. Automating Human Tutor-Style Programming Feedback: Leveraging GPT-4 Tutor Model for Hint Generation and GPT-3.5 Student Model for Hint Validation, in: Proceedings of the International Learning Analytics and Knowledge Conference (LAK), pp. 12–23.
- Prasad, P., Sane, A., 2024. A Self-Regulated Learning Framework using Generative AI and its Application in CS Educational Intervention Design, in: Proceedings of the Technical Symposium on Computer Science Education (SIGCSE), pp. 1070–1076.
- Prather, J., Denny, P., Leinonen, J., Becker, B.A., Albluwi, I., Craig, M., Keuning, H., Kiesler, N., Kohn, T., Luxton-Reilly, A., MacNeil, S., Petersen, A., Pettit, R., Reeves, B.N., Savelka, J., 2023. The Robots Are Here: Navigating the Generative AI Revolution in Computing Education, in: Proceedings of the Working Group Reports on Innovation and Technology in Computer Science Education (ITiCSE-WGR), pp. 108–159.
- Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., Finn, C., 2023. Direct preference optimization: Your language model is secretly a reward model. arXiv preprint arXiv:2305.18290 .
- Rum, S.N.M., Ismail, M.A., 2017. Metocognitive support accelerates computer assisted learning for novice programmers. *Journal of Educational Technology & Society* 20, 170–181.
- Ryan, M., Ryan, M., 2014. The pedagogical balancing act: Teaching reflection in higher education. *Teaching in Higher Education* 19, 144–155.

- Sankaranarayanan, S., Kandimalla, S.R., Bogart, C.A., Murray, R.C., Hilton, M., Sakr, M.F., Rosé, C.P., 2022. Collaborative programming for work-relevant learning: Comparing programming practice with example-based reflection for student learning and transfer task performance. *IEEE Transactions on Learning Technologies* 15, 594–604.
- Schön, D.A., 2017. *The Reflective Practitioner: How Professionals Think in Action*. volume 73.
- Seufert, T., Schüler, A., Lehmann, J., 2024. The role of cognitive load in self-regulated learning. *Educational Psychology Review* 36, 441–465.
- Shen, Y., Ai, X., Raj, A.G.S., John, R.J.L., Syamkumar, M., 2024. Implications of ChatGPT for Data Science Education, in: *Proceedings of the Technical Symposium on Computer Science Education (SIGCSE)*, pp. 1230–1236.
- Singh, A., Brooks, C., Wang, X., Li, W., Kim, J., Wilson, D., 2024. Bridging Learnersourcing and AI: Exploring the Dynamics of Student-AI Collaborative Feedback Generation, in: *Proceedings of the Learning Analytics and Knowledge Conference (LAK)*, pp. 742–748.
- Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.F., 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems* 33, 3008–3021.
- Student, 1908. *The Probable Error of a Mean*. *Biometrika* .
- Taguma, M., Barrera, M., 2019. *Oecd future of education and skills 2030: Curriculum analysis*. Dispon. Su <https://www.oecd.org/education/2030-Project-Learn/>. Pdf .
- Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A.E., Sarkar, A., Sellen, A., Rintel, S., 2024. The metacognitive demands and opportunities of generative ai, in: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–24.
- Team, J., 2018. JupyterLab. <https://github.com/jupyterlab/jupyterlab>.

- Valova, I., Mladenova, T., Kanev, G., 2024. Students' Perception of Chat-GPT Usage in Education. *International Journal of Advanced Computer Science & Applications (IJACDA)* 15.
- Vygotsky, L.S., Cole, M., 1978. *Mind in society: Development of higher psychological processes*. Harvard university press.
- Wang, S., Mitchell, J.C., Piech, C., 2024. A Large Scale RCT on Effective Error Messages in CS1, in: *Proceedings of the Technical Symposium on Computer Science Education (SIGCSE)*, pp. 1395–1401.
- Wang, Y.H., Wu, S.Y., Wang, H.H., 2017. Effects of guided reflection on positive thinking, self-satisfaction, and appreciation among college students. *Journal of Nursing Research* 25, 192–200.
- Warshauer, H.K., 2015a. Productive struggle in middle school mathematics classrooms. *Journal of Mathematics Teacher Education* 18, 375–400.
- Warshauer, H.K., 2015b. Strategies to support productive struggle. *Mathematics teaching in the middle school* 20, 390–393.
- Wirth, J., 2009. Promoting self-regulated learning through prompts. *European Journal of Psychology of Education* 24, 91–104.
- Yilmaz, R., Yilmaz, F.G.K., 2023. The Effect of Generative Artificial Intelligence (AI)-based Tool Use on Students' Computational Thinking Skills, Programming Self-Efficacy and Motivation. *Computers and Education: Artificial Intelligence (CAEAI)* 4.
- Zamfirescu-Pereira, J., Qi, L., Hartmann, B., DeNero, J., Norouzi, N., 2023. Conversational Programming with LLM-Powered Interactive Support in an Introductory Computer Science Course. *NeurIPS'23 Workshop on Generative AI for Education (GAIED)* .
- Zamfirescu-Pereira, J., Qi, L., Hartmann, B., DeNero, J., Norouzi, N., 2025. 61A Bot Report: AI Assistants in CS1 Save Students Homework Time and Reduce Demands on Staff.(Now What?), in: *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, pp. 1309–1315.

- Zarestky, J., Bigler, M., Brazile, M., Lopes, T., Bangerth, W., 2022. Reflective writing supports metacognition and self-regulation in graduate computational science and engineering. *Comput. Educ. Open* 3, 100085.
- Zhan, Y., Yan, Z., 2025. Students' engagement with chatgpt feedback: Implications for student feedback literacy in the context of generative artificial intelligence. *Assessment & Evaluation in Higher Education* , 1–14.
- Zimmerman, B.J., 2002. Becoming a Self-regulated Learner: An Overview. *Theory into practice* 41.
- Zimmerman, B.J., 2011. Motivating self-regulated problem solvers. *The Nature of Problem Solving* , 203–228.