Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025 ASEE
Paper ID #47473

Help or Hype? Exploring LLM-based Chatbots in Self-Regulated Learning

Ryan Tsang, University of California, Davis

Ryan Tsang is currently a PhD Candidate in Electrical and Computer Engineering and is interested in Embedded Systems, Firmware Security, Engineering Education and Evidence-based Teaching Practices.

SYDNEY Y WOOD, University of California, Davis

Help or Hype? Exploring LLM-based Chatbots in Self-Regulated Learning Tasks

Tsang, Ryan
University of California, Davis
Wood, Sydney
University of California, Davis
Kumar, Sarbani
University of California, Davis
Homayoun, Houman
University of California, Davis

Abstract

In this Empirical Research Full Paper, we explore the effects of chatbot usage on student performance in self-regulated learning tasks conducted in a classroom setting. The increasing use of generative artificial intelligence (AI) and large language models (LLMs) in STEM education have resulted in thought-provoking conversations regarding its potential benefits and dangers. While sophisticated LLM-based chatbots developed for pedagogical purposes (i.e., context-aware information retrieval, conversational feedback, problem-solving, etc.) may offer unprecedented accessibility and efficiency in multidisciplinary subjects, they also threaten academic integrity and rigor through abuse or hallucination. In this exploratory study, we attempt to determine the effects of chatbot usage on student learning in the context of an upper-division embedded systems lab. We designed five self-regulated learning tasks—completed by students (N=49 of 60) at the beginning of each lab module—each including a short assessment. We then employed a pseudo-random counterbalanced longitudinal design on four of the five tasks, where students used LLM-based chatbots to prepare for half of their assessments. In the fifth task we re-randomized participation groups for a standalone experiment with different motivational conditions. These experiments attempted to measure the effects of chatbot use on short-term performance of students' comprehension and problem-solving. We report experimental results for the longitudinal design, as well as the standalone design and discuss our observations. In addition, we present students' self-reported utilization strategies and sentiments regarding their use of chatbots in preparation for the assessments alongside our own analyses of their chatlogs to compare and contrast students' perceptions and their actual interaction patterns. We note from the longitudinal study that, contrary to students' generally positive attitude toward it, the use of LLM-based chatbots did not appear to have any predictive power on performance outcomes. Finally, we call for continued empirical research on the efficacy of LLM-based technologies in STEM education and propose future research directions in exploring their impact on teaching and learning.

1 Introduction

The introduction of OpenAI's ChatGPT in November 2022 [1] triggered an unprecedented surge of interest in applications of artificial intelligence (AI) based on Large Language Models (LLMs) and their underlying transformer architecture.

In particular, LLMs appear to be exceptional in applications that involve human interaction, information retrieval, and summation, making them an attractive prospect for improving the effectiveness and accessibility of education in the digital age [2, 3, 4]. However, the teaching community has raised substantial concerns regarding academic integrity, student learning, ethical application,

and the dynamics of human-AI interaction [4, 5, 6, 7, 8]. While empirical studies on LLM usage in education have been conducted in this early stage of adoption, given the current novelty of LLMs in education and the myriad ways they might be incorporated into an educational setting, additional research is crucial for better understanding the short-term and long-term effects of LLM-based AI on teaching and learning in computer science.

Due to the relative lack of evidence from early research in this area, we believe the immediate effects of using generative AI in classroom contexts remain unclear, leading to the research questions that motivate this exploratory study:

RQ1: What are the short-term effects of utilizing LLM chatbots to assist in self-regulated learning tasks on student performance?

RQ2: What strategies do students employ when using LLM chatbots in their self-study?

RQ3: What are student attitudes towards the use of LLM chatbots in their self-study?

In this paper, we present an exploration of the effects of LLM-based chatbots like ChatGPT on learning outcomes by assessing student performance on in-person formative assessments in a series of self-regulated learning tasks. The study was conducted during the winter quarter of 2024 at a large public research university in the context of an upper-division introductory embedded systems course for electrical engineering, computer engineering, and computer science majors. In addition, we collect and present survey data gauging student sentiments on their use of LLM-based chatbots, as well as initial observations on the chatlogs collected from students.

2 Background & Related Work

Self-regulated learning (SRL) is broadly defined as the ways in which individuals regulate their own cognitive processes within an educational setting. In the expansive psychological science literature on the topic, SRL theories generally categorize students' cognitive and behavioral processes across several phases that typically include: preparation, characterized by behaviors associated with goal-setting; performance, characterized by execution and monitoring of goal-directed tasks; and appraisal, characterized by reflection and adaptation [9, 10]. Research across multiple academic domains suggests that SRL-based interventions can improve student learning outcomes at various educational levels [10, 11, 12, 13, 14, 15, 9, 16, 17, 18, 19].

Several major models of SRL have emerged that have seen some justification from empirical studies in educational psychology, including Zimmerman's Cyclical Phases model; Boekaerts' Dual Processing model; Winne and Hadwin; Pintrich; Efklides; Hadwin, Järvelä, and Miller [10].

We designed self-study tasks aimed at promoting students' agency in learning course material. Due to SRL's evident effectiveness, we adopt the SRL framework to evaluate students' use of LLMs during the self-study task. We aimed to answer our research questions in the context of these tasks by designing an experiment in which task instructions were modified to permit or prohibit the use of generative AI during their completion. To answer **RQ2** in particular, we categorize the assigned task and the chatbot usage data using the Winne & Hadwin model of SRL [20].

2.1 The Winne & Hadwin Model

The Winne & Hadwin (W&H) model adopts the perspective of studying and learning as information processing tasks and proposes four basic, weakly recursive phases of learning: (1) task definition, (2) goal setting and planning, (3) enacting study tactics and strategies, and (4) metacognitive adaptation in studying behaviors [20, 21]. Furthermore, they posit that each of these phases can be analyzed along 5 dimensions: *conditions*, *operations*, *products*, *evaluations*, and *standards*—whose interactions and relationships are described in what they call the COPES model of study tasks [20]. Greene and Azevedo [21] conducted a theoretical review of the model and noted 113 studies that provide empirical support for various aspects of the model. The W&H model is also used as a framework for interpreting trace data collected from students' study activities to measure SRL, making it a practical choice for interpreting experimental results [22, 14].

We used the W&H model to frame our experiment's SRL tasks by mapping the subtasks to the different phases of learning and categorizing them using the COPES model. This allowed us to better describe the operations involving generative AI by leveraging the additional context the SRL model provides. We are primarily using W&H as a meaningful framework for describing and discussing our observations, as well as identifying when students employ generative AI and how that affects task completion. The W&H model's focus on concrete information processing maps to our study's task better than other foundational SRL theories, which place more emphasis on internal psychological states that our study did not measure. We considered adopting other SRL models specific to computer science education, such as Loksa & Ko's stages of programming problem solving [18] and Prasad & Sane's proposed SRL model integrating generative AI [23]. However, we concluded that Loksa & Ko's model was too specific to the context of problem-solving tasks, and Prasad & Sane's model, given its publication after our data was collected, would be inappropriate to retrofit onto our experiment.

2.2 Related Work

The vast majority of existing work related to LLMs in education focused on students' and teachers' internal states and sentiments in relationship to the technology, case studies on curriculum integration, tool development, and position papers [24, 25, 23, 4, 6, 7, 26, 27]. Given that comprehensive LLMs became commercially available within the last three years and that designing and conducting studies on LLMs use in the classroom is very challenging, the pool of empirical studies investigating the effects of LLMs on learning outcomes is predictably limited. Nonetheless, there are a still a number of such studies we identified across various domains that investigate LLMs' effects on learning [28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38].

Experimental studies on LLM use in higher education are incredibly difficult to design and execute, because they rely on the students choice to use LLM even if LLMs are available for them to use, resulting in selection bias and missing data [33]. In addition, classroom usually employ multiple learning strategies and practices at the same time, which makes isolating the effect of the LLM difficult to parse from other learning practices [38, 30]. Therefore, some studies report on perceived benefits or observational description of students experience with LLMs without statistical evidence, or without clear definitions of learning outcomes [28, 32]. While studies report a benefit to student performance on the immediate tasks in which LLMs were used [34, 35, 37, 36], thus far, no comprehensive evidence suggest that these benefits extend to

summative assessments or final grades [34, 35, 37, 30, 31]. In fact, some studies suggest that certain modes of LLM usage may have detrimental effects on summative assessment performance post-intervention, in which access to LLMs is removed [36, 30, 34]. Hence, the majority of studies reporting benefits of LLMs focus more on student engagement, interaction patterns, and behaviors, or student perceptions, such as satisfaction, perceived benefit, self-efficacy, or motivation [33, 37, 25, 39, 40, 41, 26, 42, 7, 43, 27, 44, 45, 46].

We discuss the relevance of this work at further length in Section 5 but note here that our study differs significantly in context, as our tasks are not assessing programming ability specifically, but broader knowledge and problem-solving skills related to computer engineering and embedded systems.

3 Methods

To test the potential impact of LLMs in SRL, we designed a 2-stage study consisting of a counterbalanced repeated measures experiment, and a randomized controlled trial using an identical measurement format.

The contents of this sections primarily describe the first stage of the study, as the design of stage 2 was nearly identical to that of stage 1. The changes made in stage 2 are outlined in Section 3.4.Our research methodology was reviewed and approved by our university institutional review board (IRB) and all data analyzed and published in this paper was obtained with the informed consent of participants.

3.1 Participants & Course Context

Participant Characteristics: Participants were students recruited from a 60-person upper-division course on introductory embedded systems. Forty-nine students consented to the use of their data for research purposes. Demographic data was collected to assess the characteristics of the sample and to assess potential confounds in the two counter-balancing groups. The majority of participants were juniors or seniors in computer science, computer engineering, or electrical engineering. All participants were in good academic standing. The majority of students reported a GPA between 3.4 and 3.79. Of the 49 students in our sample, 35% spoke English as a second language, 16% were transfer students, 18% were first-generation college students, 22% were international students, and 18% did not identify as a man. Comparing the two counterbalancing groups the only significant difference is in class composition ($p < .01$). There were more seniors in the group that used AI on Checkpoints 1 and 3 than the group that used AI on Checkpoints 2 and 4. See Appendix A for full demographic breakdown.

Course Description: We collected data in an upper-division computer science course at a large public research-intensive university. The course is designed to be a thorough introduction to the core concepts of embedded systems (i.e.,input/output, memory mapping, wired communication protocols, interrupts , etc.). The majority of coursework is associated with lab modules, which comprised roughly (35%) of their final course grade. Students are expected to complete the modules both in and out of the scheduled lab sections. Each module had milestones comprised of several implementation steps followed by a post-lab assignment. The course contained a total of four modules of varying degrees of difficulty; each module spanned four in-person sessions, ex-

cept for Module 1, which only spanned two. All students completed labs in dyads. Students chose their own partners. Dyads were generally stable for the duration of the quarter, barring outstanding circumstances.

Course Statement on Chatbots: To provide students with additional structure in their chatbot usage, we provided a brief statement on allowed, disallowed, and recommended usage patterns at the beginning of the course. In addition, during the preparation phase of the assessment tasks, teaching assistants instructed intervention group students to use chatbots as exploratory tools to gain quick familiarity with topic terminology, but then refine their understanding using conventional research methods. However, these instructions were not strictly enforced. Any strategy was permissible so long as it did not violate academic integrity and outlined allowable use principles. We note that students were allowed to use copy-and-paste functionality while working on assessment tasks during the preparation phase.

3.2 Experimental Design & Procedure

We adopted a pseudo-random counterbalanced longitudinal design for stage 1 of our study. On the first day of each Lab Module, students were instructed to complete an SRL task that consisted of a 1.5-hour preparation phase and a 20-minute assessment phase (Checkpoint). We assigned students based on lab section to one of two order conditions. Group 1 was instructed to use AI during the preparation phase of Modules 1 and 3, and barred from AI use in the preparation phase of Modules 2 and 4. Group 2 had the opposite order: barred from AI use in Modules 1 and 3, but instructed to use it in Modules 2 and 4. The dyads in Section 1 were assigned to the first order condition (Group 1), and the dyads in Section 2 were assigned to the second order condition (Group 2). The dyads in Section 3 tested out a true experimental design; they were randomly assigned to either Group 1 or 2, and seated such that physical barriers in the room divided them. In-person teaching assistants were trained to conduct the experiment and ensure that intervention conditions were not violated.

SRL Task Phases: The SRL tasks can be broken down into their chronological phases and physical components and mapped to the W&H SRL model. The preparation phase of our SRL task format may in fact encompass all 4 stages of learning: (1) *task definition*, in which students are given the task instructions and intended learning outcomes that set the external conditions of the task; (2) *goal setting and planning*, in which students implicitly or explicitly decide on a strategy for meeting those learning outcomes; (3) *enacting study tactics and strategies*, in which students carry out the operations of their strategy and generate products in the form of well-cited, digitally compiled notes; and (4) *metacognitively adapting studying* as they iteratively assess their notes and knowledge relative to internally and externally-defined task conditions and standards. The assessment phase constitutes a final, externally imposed iteration on stage 3, in which the students' completed checkpoint assessments serve a final product from which they will receive an evaluation. This takes the form of solutions to the checkpoint assessments which are later released to students and which serve as a final standard against which products were monitored. Students may engage in another phase of metacognitive adaptation at this point before repeating the cycle in preparation for mini-exams that cover the same topics.

The presence of an LLM-based chatbot during the preparation phase is thus indicated in the task definition phase and depends on the experimental group the subject belongs to (phase 1, condi-

tions). Moreover, the LLM has the potential to influence each of the COPES facets at any phases of the SRL process. For example, it might assist in creating a study strategy (phase 2, operations/products), information search and synthesis (phase 3, operations/products), or even feedback on current understanding (phase 3/4, evaluations/standards).

SRL Task Timeline

During the preparation phase, teaching assistants gave student dyads a brief overview of topics related to the corresponding lab module and instructed students to study these topics further based on a set of learning objectives included in the assessment task description. Students were instructed to take well-cited digital notes for use during the checkpoint assessment and later submission.

During the assessment phase, students individually completed a checkpoint assessment. Students were allowed to use only their digital notes as reference—further interaction with chatbots, search engines, or other sources was disallowed. Dyad members were allowed to communicate silently with one another but not with members of other dyads. Post-assessment surveys were administered as part of the students' post-lab assignments at the end of each module and were collected via web form.

Measures: We offered a demographic survey at the beginning of the instructional quarter. The survey was offered for extra credit and contained the consent form for the study.

Checkpoint assessment results were collected and graded using a 3-pass process. Graders conducted a first pass to familiarize themselves with questions and the range of answers, then a second pass to create detailed rubric items based on answer classifications, ending with a final pass with the finalized rubric to ensure all students were scored by the same metrics. Rubrics were created on a per-question basis and credit was assigned based on answers' resemblance to solution keys and demonstrated understanding of the questions' underlying concepts.

Post-assessment surveys asked students to report the perceived usefulness of the chatbot(s) they used during preparation and prompted an open-ended reflection on how they used the chatbot(s) during preparation. Their reflections were manually labeled based on the metacognitive strategies they reported employing. They contained 4 questions of interest:

Q1: Did you use Generative AI during self-study for this lab? (Yes/No)

Q2: How helpful did you find using GenAI to prepare for the checkpoint assessment for this lab? (Likert Scale 1-5)

Q3: What GenAI chatbots/engines did you use if any? (Multiple Select)

Q4: Describe how you used ChatGPT to facilitate your self-study.(Open-Ended)

3.3 Materials

Checkpoint assessment topics varied based on the contents of each module. Assessments each contained two classes of questions categorized by Bloom's Taxonomy:

Knowledge/Comprehension: questions meant to assess basic understanding of the introduced concepts (easier).

Application/Analysis: questions requiring problem-solving, design, or debugging, which assess the depth of understanding of the introduced concepts (more difficult).

Before administration, subject matter experts reviewed the checkpoint assessment questions to increase our confidence that each class of questions was at the appropriate level of difficulty.

3.4 Stage 2 Modifications

Our randomized control trial occurred within the same course context as stage 1, with the same pool of participants, using an additional 5th checkpoint that was not associated with any existing lab module. However, instead of retaining the same experimental groups, we randomly assigned dyads within each section to produce new experimental and control groups that spanned all three lab sections. To gauge the potential effects of student motivation, we altered the incentive structure of the task by rewarding extra credit points proportional to their score on the checkpoint assessment. Checkpoint 5 contained the same two classes of questions as the other checkpoints. Aside from these alterations, the measurement was conducted identically to Stage 1.

3.5 Analyses

Chatbot Log Analysis: We performed qualitative coding to assess the strategies that students used in querying chatbots. We used a predefined classification scheme in labeling the chatbot data. Specifically, we looked at whether students utilized AI to enact higher levels of self-regulated learning. The labels we used are described in Table 3. We had a coder read through all the data twice to ensure that all queries were accurately labeled.

The goal of this data is to describe the AI search strategies students use during self-regulated learning. Therefore, we performed comprehensive descriptive statistics on the label data and did not use any inferential statistics, as we were not making any comparisons between groups.

Assessment Analysis: For Stage 1, we employed a Hierarchical Linear Modeling (HLM) framework [47, 48] to analyze the impact of the AI's usage on checkpoint accuracy. HLM simultaneously estimates between-group differences and within-group changes over time by clustering variances at multiple grouping levels. This method parses apart individual differences and group membership to isolate the effects of the experimental condition. The models were fit using Full Maximum Likelihood to account for missing observations. We used the `lmer` package in R to analyze the data. Due to the nature of the data collection procedures, we preregistered analyses with multiple nested levels. Specifically, scores were nested by checkpoint number, student, dyad, group, and section. However, the sample size that we collected only had sufficient power to fit a model nested by student and group. Due to ethical constraints in obtaining consent, many dyads only contained a single individual and, thus, posed a significant convergence issue.

For Stage 2, we randomized each dyad to either AI-use or no AI-use in preparation for the assessments. To analyze this checkpoint data we use a Welch's t-test to assess the difference in performance across the two groups.

Post Assessments: We performed a general qualitative analysis on *Q4* of the post-assessment survey to label responses for further analysis. Labeling was performed in a 3-pass process, in which labels were dynamically defined during the first pass by multiple encoders, and reapplied during

Table 1: Number of User Queries to ChatGPT

Checkpoint #	Mean	*SD*	median	IQR
1	6.23	5.9	5	7.75
2	7.25	5.42	5	1.75
3	4.25	2.82	3	4
4	3.8	1.61	3	2
5	3.92	1.66	4	1

Table 2: Overall Label Frequency Across Chatlogs

Label	Mean	*SD*	Median	IQR
Original	17.92	25.34	0	33.33
Follow-up	5.21	10.95	0	0
Reworded	33.45	36.62	20	60
Copied	40.81	41.88	25	100
Irrelevant	1.38	7.82	0	0
Cheating	1.22	11.04	0	0

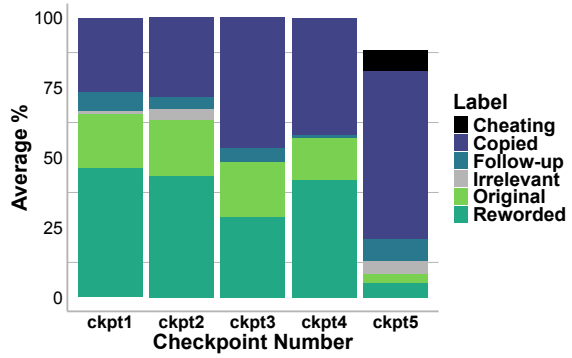


Figure 1: Average Percent of Label per Chatlog by Checkpoint

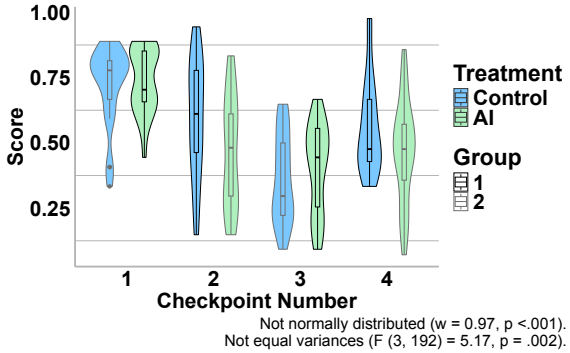


Figure 2: Assessment Score Distributions by Group and Checkpoint

the second pass to ensure a consistent set of labels. The set of labels was then reduced to remove redundancies and renamed labels were reapplied to the data. Label definitions did not occur according to any preexisting classification schema, as the intent was to precisely capture themes in participant responses without bias. Responses were tagged with multiple labels if applicable. Participants only completed this survey during the labs where they were allowed to use AI, so the analysis of this data is purely descriptive.

More information on the pre-registered data collection and analysis plans can be found on our OSF registration.

4 Results

4.1 Chatlogs Results

The descriptive statistics for students' number of queries to ChatGPT show relatively low engagement with the tool (see 1). Each query was labeled in every chatlog using the categories described in Table 3. Table 2 and Figure 1 show the average percentage of each label in the chatlogs overall and by checkpoint, respectively. We use individual chatlogs as the unit of analysis rather than individual students, because students are in dyads, so one file is submitted for two students. Checkpoint 1 and 2 represent time-point one for their respective order conditions in Stage 1, whereas Checkpoints 3 and 4 are time-point two. The general trend that we see is that from time-point 1 to time-point 2 students move away from developing prompts for the AI (Reworded, Original, and Follow-up) and instead rely heavily on copying and pasting the learning and checkpoint objectives. In addition, they prompt AI less frequently in time-point 2 compared to time-point 1. Checkpoint

Table 3: Description of Chatlog Labeling Categories

Category	Description	Example
Original	Student created an original AI query related to course topic	When the C code is translated into assembly, is it usually in x86? Or is it in some other language for embedded systems?
Follow-up	Student AI query asks for elaboration on a previous AI response	
Reworded	Student's AI query is reworded from the learning objectives, Checkpoint Objectives or Lab Instructions	What are build and flash process for microcontrollers?
Copied	Student's AI query was directly copied from Learning Objectives, Checkpoint Objectives, or Lab instructions	Know build and flash process for microcontrollers
Irrelevant	Student's AI query was about formatting, communication style, or something unrelated to the class	Respond in a conversational New York style accent. "whaddya mean?"
Cheating	Student's AI query pertained the checkpoint assessment questions (which they were not allowed to use AI to answer)	Directly copied checkpoint assessment questions

Table 4: Model Summary

Model Fit

AIC= -98.69 *BIC*= -62.86
Pseudo-R² (fixed effects) = 0.35
Pseudo-R² (total) = 0.63

Random Effects

Group	Variance	Std. Dev.
student:group	.009	0.07
student	.005	0.09
Residual	.019	0.15

Grouping Variables

N	ICC
49	0.27
49	0.16

Fixed Effects

Group	Est.	Std. Err.	*t*-value	D.F.	*p*-value[2]
β_0:Control C1 (Grp. 2)	0.72	0.04	20.05	120.11	.00
β_1:Control C2 (Grp. 1)	-0.11	0.05	-2.15	120.11	.03
β_2:Control C3 (Grp. 2)	-0.36	0.04	-9.36	137.04	.00
β_3:Control C4 (Grp. 1)	-0.17	0.05	-3.16	124.188	.00
β_4:Intervention C1 (Grp. 1)	0.01	0.05	0.22	122.13	.83
β_5:Intervention C2 (Grp. 2)	-0.16	0.09	-1.85	71.83	.07
β_6:Intervention C3 (Grp. 1)	0.03	0.06	0.62	137.87	.54
β_7:Intervention C4 (Grp. 2)	-0.10	0.09	-1.14	73.13	.26

[2] *p*-values calculated using Satterthwaite D.F.

5 (Stage 2) is standalone because across all lab sections dyads were randomly assigned to AI-use conditions. Both trends persisted in Checkpoint 5.

4.2 Assessment Results

Stage 1 Assesmment Results. The distributions of checkpoint assessment scores by AI use and group are charted in Figure 2. We fit a mixed-effects hierarchical linear model to assess the impact of AI use on students' checkpoint assessment scores. Table 4 shows the results of the model. The intercept is the average score for the control group at Checkpoint 1 (Group 2), $\beta_0 = 0.72$. The next three coefficients ($\beta_1 = -0.11, \beta_2 = -0.36, \beta_3 = -0.17$) show the expected change in average score for the control group at each checkpoint compared to Checkpoint 1. All three coefficients show a significant decrease in average score compared to Checkpoint 1, $p < .05$. It's important to note that even though comparisons were made across groups, the nested structure controls for the effect of the shared variance within-person and within-groups.

In reference to the effects of AI use, $\beta_4 = 0.01$ represents the change in average score for the AI use group at Checkpoint 1 (Group 1) compared to the control at Checkpoint 1 (Group 2). We found no evidence that using AI influenced students' scores at Checkpoint 1, $p > .05$. The last three

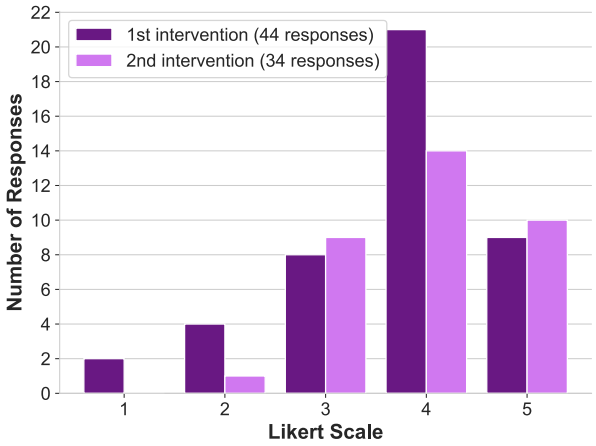


Figure 3: Reported Usefulness of Chatbots

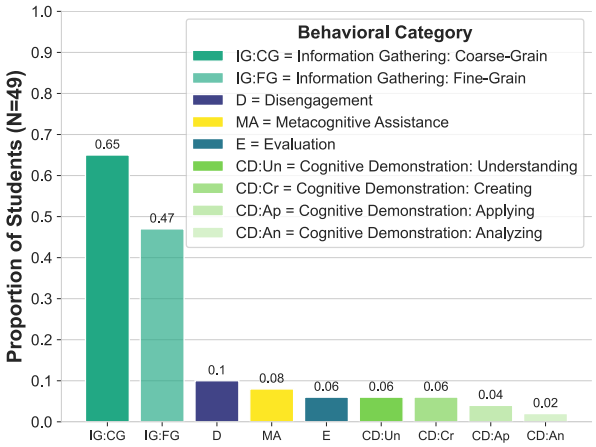


Figure 4: Behavioral Categorization of Responses by Label

coefficients ($\beta_5 = -0.16, \beta_6 = -0.03, \beta_7 = -0.10$) represent the change in average score of the intervention group compared to the control group at that time point. At each checkpoint, we found no significant differences in checkpoint score for any checkpoint, $p > .05$. The random effects and ICC suggest only weak clustering of variances at the student and group level.

Stage 2 Assessment Results. For Checkpoint 5, we saw no statistically significant differences in assessment performance between the AI-use ($M = 78.16, SD = 12.91$) and No AI-use ($M = 74.36, SD = 14.76$) groups, $t(42.67) = 0.92, p > .1$.

4.3 Survey Results

Survey question responses for *Q2* are charted in Figure 3. It reports the distribution of student responses across a Likert scale of 1 to 5, where a higher score indicated a higher degree of perceived usefulness during assessment preparation. There is an average reported score of 3.70 during the first intervention and of 3.97 during the second intervention; however, a repeated-measures analysis showed no statistical difference in the reported helpfulness from the first use to the second use, $t(33.9) = 33.90, p > .05$.

As *Q1* of the survey was used primarily as a filtering question, its responses are not charted. In response to *Q3*, nearly all participants reported relying primarily upon ChatGPT; only participants in 2 dyads reported using Google Bard instead of or in addition to ChatGPT.

A small number of students also used their response to *Q4* to comment on their perceptions of ChatGPT's usefulness; these were both positive and negative.

We have defined a set of behavioral categories to describe how students utilize chatbots. These behavior categories—defined in Table 5—were synthesized from a set of qualitative intermediate labels that captured a variety of specific usage strategies. Intermediate labels can be found in our OSF repository. Figure 4 shows the proportion of students who reported each behavioral category based on responses to *Q4*.

As shown in Figure 4, students reported using chatbots primarily for Information Gathering (*IG:CG* = 65% & *IG:FG* = 47%). Overall, 24.49% of students reported using chatbots to enhance their

Table 5: Label Categories

ID	Category	Description
CD	*Cognitive Demonstration*	chatbot asked to assist user by demonstrating a cognitive task
CD:An	↪ *Analyzing*	↪ chatbot asked to perform an *analysis* task
CD:Ap	↪ *Applying*	↪ chatbot asked to perform an *applying* task
CD:Cr	↪ *Creating*	↪ chatbot asked to perform a *creating* task
CD:Un	↪ *Understanding*	↪ chatbot asked to perform an *understanding* task
D	*Disengagement*	chatbot is used to sidestep or disengage from cognitive task
E	*Evaluation*	user engages in verification of the chatbot's outputs
IG	*Information Gathering*	chatbot is used to gather information
IG:CG	↪ *Coarse-Grain*	↪ gather general information
IG:FG	↪ *Fine-Grain*	↪ gather specific information
MA	*Metacognitive Assistance*	chatbot is used to assist user with metacognitive task

cognitive effort. These students reported enhancing their higher-order cognitive strategies through metacognitive assistance (*MA* = 8%), evaluation (*E* = 6%), and cognitive demonstration(*CA:Un* = 6%, CA:Cr = 6%, *CA:Ap* = 4%, *CA:An* = 2%). However, a few indicated using chatbots to avoid cognitive effort (*D* = 10%). We also looked at students who reported engaging in multiple behavioral categories (*n*=22). Through this analysis, it appeared that all but one of the students who reported avoiding cognitive effort did not also report engaging in higher-order cognitive strategies.

5 Discussion

5.1 Interpretation

The linear mixed-effects regression results found that students' performance decreased after the first checkpoint when we controlled for the within-person and within-group variances, implying that *C1* was significantly easier than the other three assessments. The estimates suggest that *C3* was the most difficult as students did worse overall on this assessment. The difficulty of *C2 & C4*, while significantly higher than *C1*, were fairly similar in performance. As lab modules cover progressively more difficult concepts, we expected that students would perform worse on the later checkpoints. However, the checkpoint questions took different forms depending on the topic tested, so it is unclear whether the difficulty stemmed from the course topics covered in the module or if the checkpoint problems themselves were more difficult.

In both data collection stages, we found no evidence that the use of AI better prepared students for the checkpoint assessments. When controlling for individual differences and group differences, and in the randomized control trial, the students in the control and the intervention groups did not perform significantly differently. Therefore, in response to **RQ1**, we did not find evidence to conclude that using chatbots during SRL has any short-term effects on performance outcomes. This observation is consistent with results from previous research, which suggested that observed benefits of LLMs are unlikely to extend to summative assessments and final grades [36, 31].

Interestingly, performance results appeared to be at odds with student survey responses. According to students' responses, participants generally reported that chatbots were more useful than other methods for preparing for the checkpoint assessment. Additionally, the high degree of copy-and-pasting observed in collected chatlogs contrasted with the relatively low levels of reported

disengagement in the survey. Since the disengagement label was synthesized from students' self-reported copy-and-pasting in a manner indicative of disengagement, this would suggest either a difference in perception among students regarding the nature of copy-and-pasting behavior, or potential issues with the accuracy of student self-reports. Students also demonstrated an insignificant increase in the average reported usefulness between the first and second interventions. However, our ability to interpret this result is limited due to the noticeable drop in response rate over time. Nonetheless, these observations may indicate a gap between students' perception of chatbot usefulness and chatbots' actual usefulness in achieving performance outcomes. This would be consistent with previous findings [31]. Therefore, in response to **RQ3**, our results indicate that although student attitudes toward chatbot utilization are positive on average, these attitudes did not translate to performance improvement.

The highest reported chatbot utilization behavior was *Coarse-Grained Information Gathering* (IG:CG). This was expected, as it is aligned with the strategy that was recommended to students at the beginning of the course. In conjunction with the results of our chatlog analysis, which indicates a high degree of copy-and-paste behavior with relatively little follow-up, this seems to imply that students simply adopt the externally recommended strategy in the *goal setting and planning* phase of SRL, then execute a particular IG:CG strategy in the *enactment* stage by copy-and-pasting the provided products from task definition (the learning objectives), allowing the chatbot to respond at will, much as they might a search engine.

Notably, *Fine-Grained Information Gathering* was also a popular behavior, which indicates that a number of students utilized chatbots to find topic-specific information, and constitutes a different learning strategy established in SRL Phase 2 and 3. By contrast, *Evaluation*—which indicated that students attempted to verify information provided by the chatbot—appeared to have a concerningly low rate of utilization. Collectively, our results may suggest that students have a tendency to rely on chatbots without fact-checking the information they receive from them, which would indicate a lack of Phase 3 and Phase 4 evaluation. However, this implication is contingent on the accuracy of student self-reporting and therefore requires further investigation. Unfortunately, only a relatively small proportion of students reported any higher-order learning strategies— such as using the chatbot to assist in *goal setting and planning* or validation and evaluation of LLM-generated products during *enactment*— but over twice as many students reported using chatbots to engage in higher-order learning strategies than those who reported disengaging behavior. However, the observed behaviors in the query are inconsistent with those self-reported behaviors.

Considered in context of the W&H phases, we place the observed strategies associated with IG and CD in phase 3, and observed MA behaviors in phase 2. Disengagement might be considered part of phase 1, as it might be indicative of initial internal conditions, such as lack of motivation or disinterest. The reported popularity of such strategies is unlikely to generalize as the instructions and AI Guidelines presented a significant threat to external validity. We did not include *Evaluation* with the above list, as it describes how a subject interacts with chatbot outputs as opposed to how the chatbot is used and is therefore irrelevant to **RQ3**.

In response to **RQ3**, we can make the limited observation that students report employing strategies characterized by *information gathering*, *cognitive demonstration*, *metacognitive assistance*, or *disengagement*, but the content of their queries suggest otherwise. We want to note that SRL and higher-order engagement with the LLM responses may occur during note-taking and peer

discussion, but we are unable to discern this from the chatlog data.

5.2 Implications

Given our interpretation of the results, it would appear that students tend to believe that chatbots are helpful in their learning without the evidence to subtantiate this belief. This may suggest an over-reliance on chatbot capabilities when taken in conjunction with our observations that at least 47% of students reported engaging in fine-grained information gathering while only 6% reported engaging in evaluation of that information.

That almost 25% of students reported engaging in higher-order cognitive strategies despite receiving no explicit instructions to do so suggests that students already have some intuition for the variety of ways that chatbots might be utilized in a constructive manner. Additionally, 10% of students reported using chatbots to avoid cognitive effort (disengage) which suggests chatbot utilization in learning can be potentially detrimental.

5.3 Threats to Validity

Our model design and demographics analysis accounted for a number of potential external variables; however, there are still limitations in our study. Students' time spent actively preparing and level of motivation to prepare for the assessment are the most significant threats to internal validity, as these variables play a role in self-regulated learning, but were not measured or controlled. Therefore, it is uncertain if these variables are evenly distributed across both groups, which may have implications on the validity of our participant selection. It has also been noted in prior work that student factors like self-efficacy, fear of failure, and prior grades can play a role in baseline usage of chatbots, which may also be a potential confound [44]. Moreover, the notable difference in graduating class between groups may have influenced these unaccounted-for variables and represent a potential confound. While allowing clustering in co-variances can help statistically control for this effect in our model, it does not provide a guarantee.

Additionally, there may be potential information bias as *C4* is different from other checkpoints in that it has one less question in the Application/Analysis category, and one of the two questions in that category contained a typo in its instructions. The typo was announced to all groups; however, it was not apparent that all students noted this change, possibly affecting the correctness of their final answer. We attempted to account for this by regrading according to solutions created for each variation of the problem with a corrected or uncorrected typo. While we cannot claim certainty, we do not believe that any information bias is occurring given the lack of any statistical significance on *C4* in our results.

6 Future

As we have acknowledged in Section 5.3, there are a number of limitations in our study that may prevent the generalization of our findings. Therefore, we propose altering our experimental design as follows: enact true randomization of the order conditions for all sections as had been piloted in lab section 3; amend the post-assessment survey to obtain student reports on motivation levels and time spent preparing for the assessment; change participation incentives to increase external motivation; and rewrite assessment quizzes for greater consistency and better scaffolding. We also

propose repeating the experiment in new course contexts to investigate whether our results are reproducible across subject domains.

Additionally, there are still potentially observable trends that we may find in the raw data. We collected participants' preparation notes and, more importantly, chat logs that have not yet undergone a thorough qualitative analysis and may yet yield additional insights in relation to performance data. In any case, such an analysis will prove helpful in understanding student habits in self-regulated learning tasks as well as their current strategies for prompting AI chatbots.

Finally, a number of additional questions arise from our observations which may provide a basis for future work:

- To what extent do students rely on information from LLM-generated responses without verification?
- Are there specific learning formats in which LLM-based chatbots can be used that might yield performance gains relative to learning outcomes?
- What are the roles of internal and/or external motivation on students' chatbot utilization strategies?
- Are there any observable long-term effects of chatbot usage students' university careers?
- How does LLM usage affect learning efficiency as opposed to quality?

7 Conclusion

In this study, we investigated the short-term impacts of LLM-based chatbot usage during self-regulated learning on subsequent assessment performance. Our exploratory study reveals that while students exhibit a positive attitude towards using LLM-based chatbots for self-regulated learning tasks, these tools do not significantly influence performance outcomes in an upper-division embedded systems lab. This could imply that students have a tendency to overestimate the usefulness of chatbot-dependent operations in the *enactment* stage of the W&H SRL model, which may lead to over-reliance on the products they provide. The contextual nature of our study limits its generalizability and necessitates further empirical research to better understand the potential benefits and limitations of LLM-based technologies in STEM education. Such research could clarify the extent and conditions under which students over-rely on chatbot information, how other variables like time and motivation influence performance, as well as what long-term effects chatbots might have on student learning and engagement throughout their academic career.

The degree of uncertainty surrounding LLMs impact on education continues to give cause for concern given its rate of adoption in higher education. In light of this, the research community should devote particular attention to continuing empirical research in this area. As this study has shown, trends derived from student self-reports may not accurately reflect the reality of behaviors or outcomes and curricular changes seeking to incorporate LLMs should be done with caution and careful observation.

References

[1] OpenAI, "Introducing ChatGPT," Nov. 2022. [Online]. Available: https://openai.com/index/chatgpt/

[2] S. Khan, "How AI Could Save (Not Destroy) Education | Sal Khan | TED," May 2023. [Online]. Available: https://www.youtube.com/watch?v=hJP5GqnTrNo

[3] S. Atlas, "ChatGPT for Higher Education and Professional Development: A Guide to Conversational AI," Jan. 2023. [Online]. Available: https://digitalcommons.uri.edu/cba%5Ffacpubs/548

[4] M. A. Cardona, R. J. Rodríguez, and K. Ishmael, "Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations," U.S. Department of Education, Office of Educational Technology, Tech. Rep., May 2023. [Online]. Available: https://tech.ed.gov/ai-future-of-teaching-and-learning/

[5] M. Amoozadeh, D. Daniels, D. Nam, A. Kumar, S. Chen, M. Hilton, S. Srinivasa Ragavan, and M. A. Alipour, "Trust in Generative AI among Students: An exploratory study," in *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, ser. SIGCSE 2024. New York, NY, USA: Association for Computing Machinery, Mar. 2024, pp. 67–73. [Online]. Available: https://dl.acm.org/doi/10.1145/3626252.3630842

[6] H. Carbonel and J.-M. Jullien, "Emerging tensions around learning with LLM-based chatbots: A CHAT approach," *Networked Learning Conference*, vol. 14, Apr. 2024. [Online]. Available: https://journals.aau.dk/index.php/nlc/article/view/8084

[7] C. K. Lo, "What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature," *Education Sciences*, vol. 13, no. 4, p. 410, Apr. 2023. [Online]. Available: https://www.mdpi.com/2227-7102/13/4/410

[8] M. Sullivan, A. Kelly, and P. McLaughlan, "ChatGPT in higher education: Considerations for academic integrity and student learning," *Journal of Applied Learning and Teaching*, vol. 6, no. 1, pp. 31–40, Mar. 2023. [Online]. Available: https://journals.sfu.ca/jalt/index.php/jalt/article/view/731

[9] M. Puustinen and L. Pulkkinen, "Models of Self-regulated Learning: A review," *Scandinavian Journal of Educational Research*, vol. 45, no. 3, pp. 269–286, Sep. 2001. [Online]. Available: https://doi.org/10.1080/00313830120074206

[10] E. Panadero, "A Review of Self-regulated Learning: Six Models and Four Directions for Research," *Frontiers in Psychology*, vol. 8, Apr. 2017. [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2017.00422/full

[11] S. Bergin, R. Reilly, and D. Traynor, "Examining the role of self-regulated learning on introductory programming performance," in *Proceedings of the First International Workshop on Computing Education Research*, ser. ICER '05. New York, NY, USA: Association for Computing Machinery, Oct. 2005, pp. 81–86. [Online]. Available: https://doi.org/10.1145/1089786.1089794

[12] N. Higgins, S. Frankland, and J. Rathner, "Self-Regulated Learning in Undergraduate Science," *International Journal of Innovation in Science and Mathematics Education*, vol. 29, no. 1, Apr. 2021. [Online]. Available: https://openjournals.library.sydney.edu.au/index.php/CAL/article/view/14804

[13] P. H. Winne and N. E. Perry, "Chapter 16 - Measuring Self-Regulated Learning," in *Handbook of Self-Regulation*, M. Boekaerts, P. R. Pintrich, and M. Zeidner, Eds. San Diego: Academic Press, Jan. 2000, pp. 531–566. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780121098902500457

[14] P. H. Winne and A. F. Hadwin, "nStudy: Tracing and Supporting Self-Regulated Learning in the Internet," in *International Handbook of Metacognition and Learning Technologies*, R. Azevedo and V. Aleven, Eds. New York, NY: Springer, 2013, pp. 293–308. [Online]. Available: https://doi.org/10.1007/978-1-4419-5546-3%5F20

[15] L. Silva, A. J. Mendes, A. Gomes, and G. F. Cavalcanti de Macêdo, "Regulation of Learning Interventions in Programming Education: A Systematic Literature Review and Guideline Proposition," in *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '21. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 647–653. [Online]. Available: https://dl.acm.org/doi/10.1145/3408877.3432363

[16] J. Prather, B. A. Becker, M. Craig, P. Denny, D. Loksa, and L. Margulieux, "What Do We Think We Think We Are Doing? Metacognition and Self-Regulation in Programming," in *Proceedings of the 2020 ACM Conference on International Computing Education Research*, ser. ICER '20. New York, NY, USA: Association for Computing Machinery, Aug. 2020, pp. 2–13. [Online]. Available: https://dl.acm.org/doi/10.1145/3372782.3406263

[17] D. Loksa, L. Margulieux, B. A. Becker, M. Craig, P. Denny, R. Pettit, and J. Prather, "Metacognition and Self-Regulation in Programming Education: Theories and Exemplars of Use," *ACM Trans. Comput. Educ.*, vol. 22, no. 4, pp. 39:1–39:31, Sep. 2022. [Online]. Available: https://dl.acm.org/doi/10.1145/3487050

[18] D. Loksa and A. J. Ko, "The Role of Self-Regulation in Programming Problem Solving Process and Success," in *Proceedings of the 2016 ACM Conference on International Computing Education Research*, ser. ICER '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 83–91. [Online]. Available: https://dl.acm.org/doi/10.1145/2960310.2960334

[19] C. Dignath and G. Büttner, "Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level," *Metacognition and Learning*, vol. 3, no. 3, pp. 231–264, Dec. 2008. [Online]. Available: https://doi.org/10.1007/s11409-008-9029-x

[20] P. Winne and A. Hadwin, "Studying as self-regulated learning," in *Metacognition in Educational Theory and Practice*, 1st ed. Lawrence Erlbaum Associates Publishers, Jan. 1998, vol. 93, pp. 277–304. [Online]. Available: https://www.researchgate.net/publication/247664651

[21] J. A. Greene and R. Azevedo, "A Theoretical Review of Winne and Hadwin's Model of Self-Regulated Learning: New Perspectives and Directions," *Review of Educational Research*, vol. 77, no. 3, pp. 334–372, Sep. 2007. [Online]. Available: https://doi.org/10.3102/003465430303953

[22] P. H. Winne, A. F. Hadwin, and C. Gress, "The learning kit project: Software tools for supporting and researching regulation of collaborative learning," *Computers in Human Behavior*, vol. 26, no. 5, pp. 787–793, Sep. 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563207001525

[23] P. Prasad and A. Sane, "A Self-Regulated Learning Framework using Generative AI and its Application in CS Educational Intervention Design," in *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, ser. SIGCSE 2024. New York, NY, USA: Association for Computing Machinery, Mar. 2024, pp. 1070–1076. [Online]. Available: https://dl.acm.org/doi/10.1145/3626252.3630828

[24] B. A. Becker, P. Denny, J. Finnie-Ansley, A. Luxton-Reilly, J. Prather, and E. A. Santos, "Programming Is Hard - Or at Least It Used to Be: Educational Opportunities and Challenges of AI Code Generation," in *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, ser. SIGCSE 2023. New York, NY, USA: Association for Computing Machinery, Mar. 2023, pp. 500–506. [Online]. Available: https://doi.org/10.1145/3545945.3569759

[25] R. Liu, C. Zenke, C. Liu, A. Holmes, P. Thornton, and D. J. Malan, "Teaching CS50 with AI: Leveraging Generative Artificial Intelligence in Computer Science Education," in *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, ser. SIGCSE 2024. New York, NY, USA: Association for Computing Machinery, Mar. 2024, pp. 750–756. [Online]. Available: https://dl.acm.org/doi/10.1145/3626252.3630938

[26] N. Raihan, M. L. Siddiq, J. C. S. Santos, and M. Zampieri, "Large Language Models in Computer Science Education: A Systematic Literature Review," Oct. 2024. [Online]. Available: http://arxiv.org/abs/2410.16349

[27] C. McGrath, A. Farazouli, and T. Cerratto-Pargman, "Generative AI chatbots in higher education: A review of an emerging research area," *Higher Education*, Aug. 2024. [Online]. Available: https://doi.org/10.1007/s10734-024-01288-w

[28] B. Qureshi, "Exploring the Use of ChatGPT as a Tool for Learning and Assessment in Undergraduate Computer Science Curriculum: Opportunities and Challenges," in *2023 9th International Conference on E-Society, e-Learning and e-Technologies*, Jun. 2023, pp. 7–13. [Online]. Available: http://arxiv.org/abs/2304.11214

[29] H. Kumar, D. M. Rothschild, D. G. Goldstein, and J. M. Hofman, "Math Education with Large Language Models: Peril or Promise?" Rochester, NY, Nov. 2023. [Online]. Available: https://papers.ssrn.com/abstract=4641653

[30] G. Jošt, V. Taneski, and S. Karakatič, "The Impact of Large Language Models on Programming Education and Student Learning Outcomes," *Applied Sciences*, vol. 14, no. 10, p. 4115, Jan. 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/10/4115

[31] T. Kosar, D. Ostojić, Y. D. Liu, and M. Mernik, "Computer Science Education in ChatGPT Era: Experiences from an Experiment in a Programming Course for Novice Programmers," *Mathematics*, vol. 12, no. 5, p. 629, Jan. 2024. [Online]. Available: https://www.mdpi.com/2227-7390/12/5/629

[32] R. Wei, K. Li, and J. Lan, "Improving Collaborative Learning Performance Based on LLM Virtual Assistant," in *2024 13th International Conference on Educational and Information Technology (ICEIT)*, Mar. 2024, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10540942/

[33] A. Nie, Y. Chandak, M. Suzara, M. Ali, J. Woodrow, M. Peng, M. Sahami, E. Brunskill, and C. Piech, "The GPT Surprise: Offering Large Language Model Chat in a Massive Coding Class Reduced Engagement

but Increased Adopters Exam Performances," *OSF Preprints*, no. qy8zd, Apr. 2024. [Online]. Available: https://ideas.repec.org//p/osf/osfxxx/qy8zd.html

[34] H. Bastani, O. Bastani, A. Sungu, H. Ge, Ö. Kabakcı, and R. Mariman, "Generative AI Can Harm Learning," Rochester, NY, Jul. 2024. [Online]. Available: https://papers.ssrn.com/abstract=4895486

[35] W. Lyu, Y. Wang, T. R. Chung, Y. Sun, and Y. Zhang, "Evaluating the Effectiveness of LLMs in Introductory Computer Science Education: A Semester-Long Field Study," in *Proceedings of the Eleventh ACM Conference on Learning Scale*, ser. LS '24. New York, NY, USA: Association for Computing Machinery, Jul. 2024, pp. 63–74. [Online]. Available: https://doi.org/10.1145/3657604.3662036

[36] M. Lehmann, P. B. Cornelius, and F. J. Sting, "AI Meets the Classroom: When Does ChatGPT Harm Learning?" Aug. 2024. [Online]. Available: http://arxiv.org/abs/2409.09047

[37] H. Kumar, I. Musabirov, M. Reza, J. Shi, X. Wang, J. J. Williams, A. Kuzminykh, and M. Liut, "Impact of Guidance and Interaction Strategies for LLM Use on Learner Performance and Perception," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW2, pp. 1–30, Nov. 2024. [Online]. Available: http://arxiv.org/abs/2310.13712

[38] I. A. Mohammed, A. Bello, and B. Ayuba, "Effect of large language models artificial intelligence chatgpt chatbot on achievement of computer education students," *Education and Information Technologies*, Jan. 2025. [Online]. Available: https://doi.org/10.1007/s10639-024-13293-8

[39] M. Liu and F. M'Hiri, "Beyond Traditional Teaching: Large Language Models as Simulated Teaching Assistants in Computer Science," in *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, ser. SIGCSE 2024. New York, NY, USA: Association for Computing Machinery, Mar. 2024, pp. 743–749. [Online]. Available: https://dl.acm.org/doi/10.1145/3626252.3630789

[40] R. Yilmaz and F. G. Karaoglan Yilmaz, "Augmented intelligence in programming learning: Examining student views on the use of ChatGPT for programming learning," *Computers in Human Behavior: Artificial Humans*, vol. 1, no. 2, p. 100005, Aug. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2949882123000051

[41] S. Rasnayaka, G. Wang, R. Shariffdeen, and G. N. Iyer, "An Empirical Study on Usage and Perceptions of LLMs in a Software Engineering Project," Jan. 2024. [Online]. Available: http://arxiv.org/abs/2401.16186

[42] S. Wang, T. Xu, H. Li, C. Zhang, J. Liang, J. Tang, P. S. Yu, and Q. Wen, "Large Language Models for Education: A Survey and Outlook," Apr. 2024. [Online]. Available: http://arxiv.org/abs/2403.18105

[43] Y. AlBadarin, M. Tukiainen, M. Saqr, and N. Pope, "A Systematic Literature Review of Empirical Research on ChatGPT in Education," Rochester, NY, Sep. 2023. [Online]. Available: https://papers.ssrn.com/abstract=4562771

[44] L. E. Margulieux, J. Prather, B. N. Reeves, B. A. Becker, G. Cetin Uzun, D. Loksa, J. Leinonen, and P. Denny, "Self-Regulation, Self-Efficacy, and Fear of Failure Interactions with How Novices Use LLMs to Solve Programming Problems," in *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*, ser. ITiCSE 2024. New York, NY, USA: Association for Computing Machinery, Jul. 2024, pp. 276–282. [Online]. Available: https://doi.org/10.1145/3649217.3653621

[45] E. D. Manley, T. Urness, A. Migunov, and M. A. Reza, "Examining Student Use of AI in CS1 and CS2," *J. Comput. Sci. Coll.*, vol. 39, no. 6, pp. 41–51, May 2024.

[46] R. Budhiraja, I. Joshi, J. S. Challa, H. D. Akolekar, and D. Kumar, ""It's not like Jarvis, but it's pretty close!" - Examining ChatGPT's Usage among Undergraduate Students in Computer Science," in *Proceedings of the 26th Australasian Computing Education Conference*, ser. ACE '24. New York, NY, USA: Association for Computing Machinery, Jan. 2024, pp. 124–133. [Online]. Available: https://dl.acm.org/doi/10.1145/3636243.3636257

[47] D. A. Hofmann, "An Overview of the Logic and Rationale of Hierarchical Linear Models," *Journal of Management*, vol. 23, no. 6, pp. 723–744, Dec. 1997. [Online]. Available: https://doi.org/10.1177/014920639702300602

[48] A. S. Bryk and S. W. Raudenbush, "Application of Hierarchical Linear Models to Assessing Change," *Psychological Bulletin*, vol. 101, no. 1, pp. 147–158, 1987.

A Participant Demographics

Table 6: Participant Demographics

Characteristic	N	Overall, N = 49[1]	group1, N = 24[1]	group2, N = 25[1]	p-value2[2]
Age	49	22.00 (21.00, 23.00)	22.00 (21.75, 23.00)	22.00 (21.00, 23.00)	0.7
Fin. Support	49	6 (6, 7)	6 (5, 7)	6 (6, 7)	0.3
e_exp	49	100 (99, 100)	100 (98, 100)	100 (99, 100)	0.8
GPA	42	3.60 (3.40, 3.79)	3.60 (3.23, 3.80)	3.60 (3.40, 3.76)	0.6
Final Grade (%)	48	91.5 (88.5, 95.2)	91.7 (88.2, 94.9)	90.8 (88.7, 95.2)	¿0.9
Mini Exams (max 64)	48	49 (46, 54)	48 (45, 54)	52 (47, 56)	0.3
Class	49				0.002
Junior		16 (33%)	3 (13%)	13 (52%)	
Senior		29 (59%)	20 (83%)	9 (36%)	
Senior+		4 (8.2%)	1 (4.2%)	3 (12%)	
ESL	48				¿0.9
No		31 (65%)	15 (65%)	16 (64%)	
Yes		17 (35%)	8 (35%)	9 (36%)	
Transfer	49				0.5
No		8 (16%)	5 (21%)	3 (12%)	
Yes		41 (84%)	19 (79%)	22 (88%)	
Gender	49				0.2
Man		40 (82%)	22 (92%)	18 (72%)	
Woman		8 (16%)	2 (8.3%)	6 (24%)	
Non-Binary		1 (2.0%)	0 (0%)	1 (4.0%)	
First Gen.	49				0.7
No		40 (82%)	19 (79%)	21 (84%)	
Yes		9 (18%)	5 (21%)	4 (16%)	
Residency	49				0.5
International		11 (22%)	4 (17%)	7 (28%)	
US		38 (78%)	20 (83%)	18 (72%)	

1: Median (Quartile 1, Quartile 3); n (%)

2: Wilcoxon rank sum test; Wilcoxon rank sum exact test; Fisher's exact test

Table 6 presents the breakdown of participant demographics and controls for variations between groups. As already noted, the only major difference between intervention groups was in the distribution of juniors and seniors.

B Post-Assessment Survey Response Labels

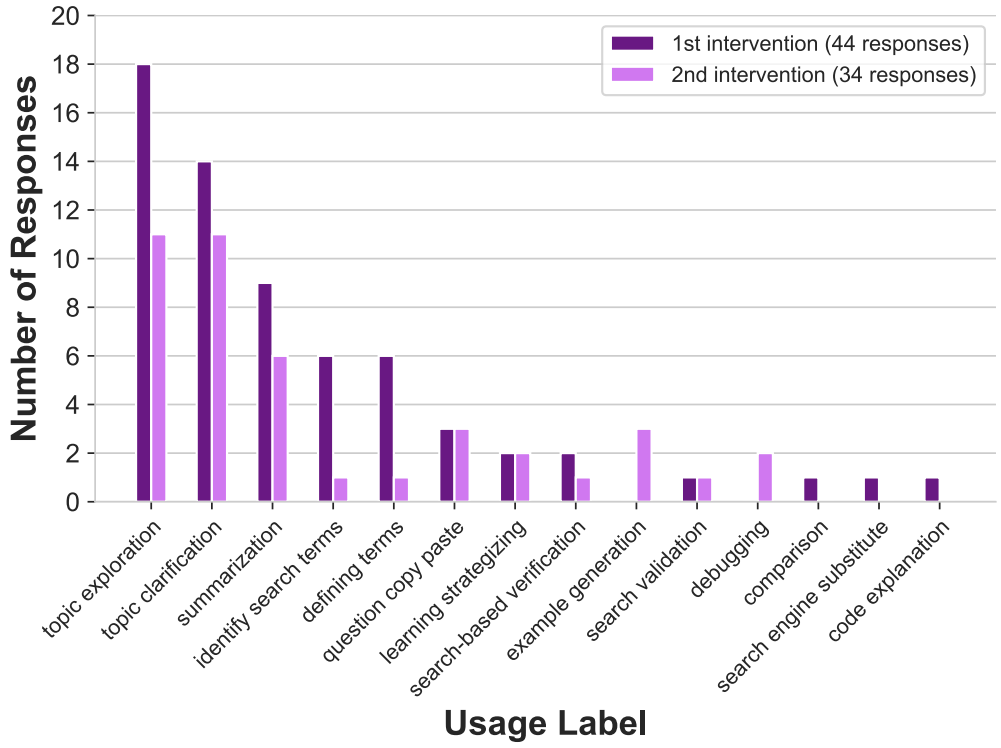
Figure 5a reports labels corresponding to the different chatbot usage strategies we were able to identify and commensurate number of responses that were identified as such. These labels are defined in Table 5b, which also contains the code indicating the behavioral category they belong to. These figures show combined data for participants' first and second interventions—i.e., first intervention group encompasses responses from Group 1 on *C1* and Group 2 on *C2* (44 total), while the second encompasses responses from Group 1 on *C3* and Group 2 on *C4* (34 total).

C Assessment Grading Principles

These were the principles adopted during rubric construction and grading:

- More specificity is better, point values will be adjusted at the end based on the level of understanding that each rubric item represents

(a) Reported Usage Strategy Frequencies



(b) Usage Label Definitions

Cat.	Usage Label	Label Description
IG:CG	*topic exploration*	chatbot was used to perform high-level exploration of topics from learning goals
IG:FG	*topic clarification*	chatbot was used to refine/clarify understanding of topic details
IG:CG	*summarization*	chatbot was used to provide summaries of requested topics
IG:CG	*identify search terms*	chatbot was used to identify topic-specific terminology for later search engine input
IG:FG	*defining terms*	chatbot was directly used to define topic-specific terminology
D	*question copy paste*	chatbot was directly given learning goals or questions copied from task instructions
MA	*learning strategizing*	chatbot was used to select strategies for learning topics
E	*search-based verification*	information provided by chatbot was later validated via search engine
CD:Cr	*example generation*	chatbot was used to generate illustrative examples of topic-specific concepts
CD:Un	*search validation*	information provided by chatbot was used as support for initial search engine research
CD:Ap	*debugging*	chatbot was used to suggest debugging strategies related to topics
CD:An	*comparison*	chatbot was used to compare concepts
IG:FG	*search engine substitute*	chatbot was used as fallback information gathering strategy when search engine yielded few/poor results
CD:Un	*code explanation*	chatbot was used to explain topic-specific code snippets

Figure 5: Usage Strategy Labels

- Partial credit awarded based on how close answer is to correct answer, unless shown work is inconsistent with answer (missing work treated separately)
- If all work is correct but a final answer is not explicitly given, award generous partial credit.
- If an answer is given that uses different terminology for tangential subjects, but understanding is obvious, apply the relevant rubric item generously.
- If the terminology was used in the question and is clearly a necessary part of understanding, apply item con-

servatively.

- If an answer was too short to discern student understanding, apply rubric items conservatively.

Generosity here is to lean towards applying rubric items to award more points, while conservative application leans towards awarding fewer points.

D Ethical Considerations

There were no direct risks posed to participants by the experimental procedures; our primary ethical considerations were regarding participant's data privacy and equity between students who opted in and opted out. To address data privacy, after the conclusion of the course, before any analysis, student data was anonymized using a randomly generated unique identifier (UID) associated with their demographic and experimental data. All personally identifiying information including name, email, and student ID were redacted from raw data before cleaning and analysis and a manual scan was performed to ensure no identities could be inferred on the basis of our cleaned data. Linking keys used to map participant data to their UID were destroyed before public release of anonymous data used for analysis. To address questions of equity, interventions were incorporated into the course structure in order to ensure all students would benefit equally from any potential gains to be had from the intervention. As extra credit was used as the motivating incentive for participation, we ensured that any credit assigned could be obtained regardless of consenting status. Moreover, in order to prevent participation status from influencing the researchers on course staff, the demographic survey results and participation status were kept confidential by an uninvolved member of the research team and shared only after final course grades were submitted. Additionally, in order to prevent advantages offered by the intervention from influencing course grade, checkpoint assessments were graded on completion rather than correctness; students were instead incentivized to take assessments seriously through the mini-exams, in which questions similar to those on the checkpoint assessments would appear.

E Open Science

Our study has been preregistered on OSF at `https://osf.io/z4juc` and our anonymized dataset and source code is publicly available at `https://github.com/comp-sotl/asee-ace-25`.