# Unlocking Insights: Investigating Student AI Tutor Interactions in a Large Introductory STEM Course


Jae-Eun Russell
University of Iowa
Iowa City, Iowa, USA
jae-russell@uiowa.edu

Anna Marie Smith
University of Iowa
Iowa City, Iowa, USA
anna-smith-1@uiowa.edu

Salim George
University of Iowa
Iowa City, Iowa, USA
salim-george@uiowa.edu

Jonah Pratt
University of Iowa
Iowa City, Iowa, USA
jonah-pratt@uiowa.edu

Brian Fodale
University of Iowa
Iowa City, Iowa, USA
brian-fodale@uiowa.edu

Cassandra Monk
University of Iowa
Iowa City, Iowa, USA
cassandra-monk@uiowa.edu

Adam Brummett
University of Iowa
Iowa City, Iowa, USA
adam-brummett@uiowa.edu



## Abstract

This study explored the use of an AI tutor and its relationship to performance outcomes in a large introductory undergraduate STEM course, where the AI tutor was integrated into the online homework system. The course included 13 weekly homework assignments, comprising 221 questions that contributed 19.5% to the final grade. Results showed that students predominantly completed homework problems without AI tutor assistance, using it selectively to address specific challenges. Patterns of AI interaction varied at both the problem and student levels, with demographic factors having little to no relationship to AI usage. Notably, the frequency of AI use was not linked to exam performance. A multi-level cluster analysis identified distinct patterns in students' use of the AI tutor during problem-solving. These patterns of use had more significant associations with performance than frequency of use alone. This paper explores these interaction patterns in depth and discusses the study's limitations and implications.



## CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Applied computing** → **Computer-assisted instruction**; **Interactive learning environments**; **E-learning**.

## Keywords

AI tutors, Higher Education, STEM





*LAK 2025, Dublin, Ireland*



ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706524

**ACM Reference Format:**
Jae-Eun Russell, Anna Marie Smith, Salim George, Jonah Pratt, Brian Fodale, Cassandra Monk, and Adam Brummett. 2025. Unlocking Insights: Investigating Student AI Tutor Interactions in a Large Introductory STEM Course. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3706468.3706524


## 1 INTRODUCTION

Intelligent tutoring systems (ITS), with the goal of providing adaptive and personalized learning experiences for students, have a long history in higher education. Numerous studies have reported benefits of ITS in higher education, including improved learning outcomes [30] [17], increased student engagement and motivation [1][31], and enhanced accessibility and flexibility [16]. However, ITS have several limitations, such as high development and operation costs [33] and more limited utility in providing assistance for open-ended or less structured problems [14].

Recent advances in generative AI present new opportunities for the evolution of ITS. The introduction of ChatGPT [22] marks a potentially transformative shift in education, but it also brings new challenges that institutions are grappling with. ChatGPT shows promise by providing students with coherent and contextually relevant responses to academic queries [21][19][2]. However, while these responses may seem appropriate, the accuracy of the content is not always reliable. This is particularly concerning for students who are just beginning to learn a topic and may lack the foundational knowledge to identify subtle inaccuracies. As a result, many educators have concerns about the inaccuracies and "hallucinations" generated by AI tools. Additionally, there are growing fears that AI could undermine academic integrity by increasing instances of plagiarism and cheating.

Amid these concerns in higher education, responses vary. Some instructors have embraced the technology, while others have chosen to ban it outright, and many remain uncertain about how to address AI in their classrooms. Since ChatGPT is widely available, students can use it at their discretion, often regardless of course

policies. Moreover, the process of detecting AI-generated content remains inconsistent, raising concerns about the potential for false accusations against students [7]. As a result, instructors face challenges in managing AI use in ways that support students' learning and ensure fairness. Given these challenges, it is crucial to examine students' use of AI to identify pedagogical strategies that mitigate its risks while maximizing its potential to enhance teaching and learning.

## 1.1 Student Attitudes toward Generative AI in Educational Settings

Several studies have investigated students' experiences with and attitudes toward generative AI tools such as ChatGPT and Microsoft Copilot. The findings generally indicate that students view AI positively, recognizing its value as a personal assistant, especially for providing immediate feedback and support in writing, coding, and various academic tasks [3]. However, students also express concerns about the reliability of AI responses and potential negative effects on learning [25]. Some studies found that certain students tend to trust ChatGPT's answers without question [6], emphasizing the need for AI literacy education to foster critical thinking when evaluating AI-provided information. In contrast, some students remain skeptical about the accuracy of AI responses, especially after receiving incorrect answers [12].

While most students have a positive attitude towards using ChatGPT in education, they also harbor ethical concerns about its use in classes and the potential negative impacts on creativity, productivity, and personal development [3] [25]. Academic integrity remains a key concern; a survey of 2,555 students revealed that over half had either used or considered using ChatGPT for academic purposes, though 70% opposed using it to write entire essays. Not surprisingly, students with greater confidence in their academic writing skills were less likely to report using it for academic purposes and were also more critical of its use by their peers. [13].

Concerns about academic integrity are prevalent among both students and educators, with a study of 389 students and 36 educators highlighting fears that AI could compromise assessments. The study recommended adapting assessments, such as presentations, to encourage critical thinking and account for AI use [28]. Adapted assessments alongside understanding and mitigating reasons students plagiarize, such as time pressure to complete their academic tasks [4] could reduce these potential threats to academic integrity. Overall, addressing these concerns, along with AI literacy education, will be essential to integrating AI responsibly in education.

## 1.2 ChatGPT and AI Tutors

Since ChatGPT's introduction [22], various studies have examined its capabilities and limitations. Results have been mixed: Sharpe et al. [27] found ChatGPT could almost perfectly solve a set of Python assignments in a computer science course, while Wang et al. [32] noted it struggled with under-specified, real-world problems due to an inability to make assumptions about missing data in the context of an engineering physics course. Another study examined ChatGPT's performance as an AI assistant using a dataset of real assessment questions from 50 courses. They found GPT-4's accuracy to be anywhere from 65.8% to 85.1% depending on which strategy was used out of eight different prompting approaches and whether they considered the majority result out of them all or not [2].

Some experimental studies have explored the impact of ChatGPT on student learning and performance. Urban et al. [29] randomly assigned students to use ChatGPT during a creative problem-solving task and found that those who used it produced more original solutions, elaborated more thoroughly, and aligned better with task goals compared to those who did not. However, in a first-year undergraduate programming course, Kosar et al. [15] found no significant differences in performance between students encouraged to use ChatGPT and those discouraged from using it, based on practical assignments and midterm exam results. These studies highlight both the potential benefits and limitations of using AI to support student learning.

Studies focusing on AI tutors have been mostly positive. Ma et al. [19] found an AI tutor in a programming course provided accurate responses 98% of the time, with most students reporting it helped their learning. Students rated generative AI tools like ChatGPT highly for their convenience, but tended to view the quality of assistance as lower than that from instructors or teaching assistants [12]. Despite these perceptions, a survey of 326 Harvard undergraduates revealed that 90% had used AI, with 25% using it instead of attending office hours or doing required readings [11].

Indeed, comparisons between AI and human tutors show promising results for AI. Students in a study by Gold and Geng [10] reported that a GPT-4-powered tutor was nearly as helpful as teaching assistants, using the AI tutor more frequently than office hours. Liu and M'hiri [18] found a virtual TA matched human TAs in accuracy while being clearer and more engaging. Nguyen et al. [21] compared student performance based on the feedback from the AI assistant (GPT-4) or the instructor in a computer science course. In three quizzes on recursion, about half of the students randomly received feedback from GPT-4, while the remaining students received feedback from the course instructor. Students in both the AI and the instructor conditions rated feedback as equally supportive. They found that the students who received feedback from GPT-4 performed better on resubmissions. GPT-4's feedback was more detailed, potentially offering better guidance for revisions.

A few studies have examined how students interact with AI tutors. Frankford et al. [8] observed two main patterns in student use of an AI tutor: (a) intensive AI use before code submissions and (b) alternating between AI and system feedback. Scholl et al. [26] found three types of interaction patterns: (a) seeking immediate solutions, (b) help with conceptual understanding, or (c) chatting extensively and engaging with AI as a study buddy.

Although existing studies provide some insights into students' attitudes toward the capabilities and impact of generative AI tutors, more exploration is needed to understand how undergraduate students interact with these tools and the effects on their learning outcomes in different contexts. Towards this end, our study is guided by the following questions related to an AI tutor that was available for assistance on homework in an introductory STEM undergraduate course:

(1) What are students' perceptions of the AI tutor, and for what purposes do they use it?

(2) What demographic and academic characteristics are associated with use of the AI tutor?
(3) Is the frequency of AI tutor use associated with student performance outcomes?
(4) Do distinct patterns of AI tutor use emerge during homework sessions? Are these patterns associated with performance outcomes?

# 2 METHODS

## 2.1 Participants and Setting

A total of 660 students (84.5%) out of 770 students enrolled in an introductory chemistry course at a large public institution located in the Midwestern USA agreed to participate in this study. Among participants, a majority (73%) were first-year students, 13% were classified as an underrepresented racial minority (URM), 26% were first-generation college students, and over half (54%) were female.

Course assessments included weekly homework, discussion participation, lab work, and exams. Weekly homework was completed using a publisher's homework platform. Each homework set consisted of 10 – 20 problems, and students were given three attempts to submit the correct answer for each problem without penalty. A total of 221 problems were assigned, accounting for 19.5% of the final grade. Within the homework platform, students had access to an AI tutor, available as a chatbot. Students could initiate an interaction with the AI tutor by typing their own message or using one of two prepopulated prompts. The AI tutor was designed to not give away the correct answer, but to instead help guide students using a Socratic approach.

## 2.2 Data

*2.2.1 Homework Interaction Data.* Students' timestamped AI interaction and problem submission data were extracted from the online homework system. Each message a student sent to the AI tutor was recorded as a chat event. Only chat events linked to assigned homework problems were included in the analysis. The final set of trace data included 227,960 problem submission events and 114,565 chat events. Problem submission events were categorized as correct (134,779 events) or incorrect (93,181 events).

Among the participants, 613 (93%) interacted with the AI tutor at least once during the semester. Among these participants, use of the AI tutor was relatively low; they interacted with the AI tutor for 14% of the problems they attempted on average. When interacting with the tutor, students sent 4.6 messages ($SD$ = 2.7) on average, indicating that once interaction was initiated, there was typically some additional dialogue between the student and AI tutor (Table 1). The frequency of interaction with the AI tutor varied across the assigned homework, ranging from 9.5% of the problems for the first homework assignment to 26.31% of the problems on the fifth homework assignment (Figure 1).

*2.2.2 Survey Data.* Participants responded to two questions related to the AI tutor. The first question asked if they had used the AI tutor at any point in the semester (Yes/No). The second question was open-ended and asked them to explain why they did or did not use the AI tutor during the semester.

*2.2.3 Demographic and Learning Outcome Data.* Gender, academic year, first-generation status, underrepresented minority status, incoming GPA ($M$ = 3.36, $SD$ =.62), homework scores ($M$ = 93%, $SD$ =7.95%), exam scores ($M$ = 65%, $SD$ = 14.19%), and final course grades were collected from institutional data after the semester concluded.

## 2.3 Clustering Technique

To understand students' interaction patterns, we selected the three homework assignments with the most AI tutor use, and implemented a multi-step clustering approach outlined in detail by Zhang et al. [34] to generate clusters at the problem and student level.

*2.3.1 Problem-Level Clustering.* Student's interactions with each homework problem were represented as a sequence of chat events and problem attempt events. Upon opening the homework problem, students could attempt to solve the problem or ask the AI tutor for help. If the student failed to solve the problem, they could either immediately make another attempt or ask the tutor for help. This process would continue until they either solved the problem, gave up, or used up all three allotted attempts.

We defined seven features meant to describe students' interactions with each homework problem as displayed in Table 2. Using the event data and the R-package bupaR [23], we were able to construct traces for student's interactions with each problem. The partition around medoids (PAM) function from the cluster R-package [20] was used to cluster problem-level traces using these seven features. The mean silhouette value [24] for clustering solutions with numbers of clusters ranging from 1 to 10 was used to determine the optimal number of clusters (*p-clusters*).

*2.3.2 Sequence-Level Clustering.* Each problem-level event trace was assigned to one of the problem clusters identified in the previous step. Therefore, students' interactions with the 44 assigned problems could be represented as a sequence of *44* problem-level clusters ($p_1, p_2, \cdots, p_{44}$). Using the TraMineR R-package [9], the dissimilarity matrix between each sequence was calculated with the optimal matching (OM) distance via the TRATE method. PAM was again used to cluster the sequences, and the number of clusters (*s-clusters*) for each sequence was determined by finding the local maximum of the average silhouette value between 2 and 10 clusters.

# 3 RESULTS

## 3.1 Students' Perceptions of the AI Tutor

Among the 660 participants, 83% reported having used the AI homework tutor at some point during the semester. Overall, most students found the AI tutor helpful. About a quarter of respondents indicated the tool was helpful but did not elaborate, while the remaining three quarters provided specific examples of how the AI tutor helped them. The most commonly cited reasons for using the AI tutor were as follows: (a) for step-by-step guidance on how to approach a problem, (b) to explain the underlying concepts related the problem, and (c) to clarify or rephrase what the problem was asking. This was not surprising given that the two prepopulated prompts were "*Start with the underlying concepts*" and "*Please rephrase the question.*" Some other reasons that were mentioned related to helping with difficult questions, helping to correct known errors, convenience/availability, checking if an answer was correct

**Table 1: Summary statistics of AI tutor use among students who utilized the tutor**

| Variable | N | Mean | Std. Dev | Min | Pctl. 25 | Pctl. 50 | Pctl. 75 | Max |
|---|---|---|---|---|---|---|---|---|
| Percent of problems with AI use | 613 | 14% | 12% | 0.45% | 4.1% | 12% | 20% | 62% |
| Number of messages | 613 | 4.6 | 2.7 | 1 | 2 | 4.6 | 6.4 | 20 |

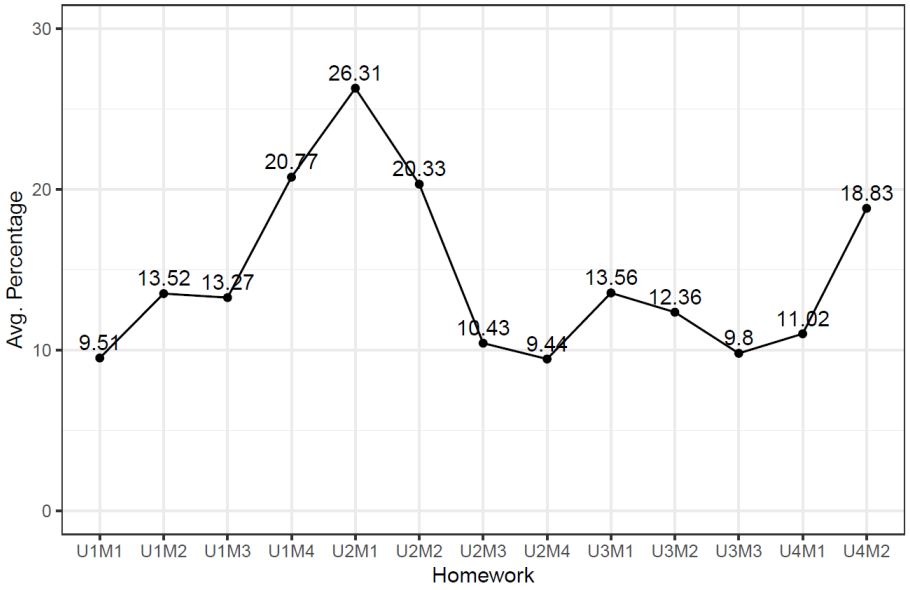


**Figure 1: Percent of problems for each assigned homework that students interacted with the AI tutor (N = 613).**

**Table 2: Features and definitions for problem-level clustering**

| Feature | Definition |
|---|---|
| Number_incorrect | Total number of incorrect attempts for a problem |
| Got_correct | If the student had a correct attempt |
| Total_attempts | Total attempts |
| Use_chat | If chat was used at all |
| Total _chats | Number of chat events |
| Incorrect_pre_chat | Number of incorrect attempts prior to the first chat event |
| First_event_chat | If 1st event for a problem was a chat event |

before submitting it, and making sure they truly understood the problem rather than guessing.

In contrast, 111 (or 17%) of respondents reported that they did not use the AI homework tutor in the course. These respondents were asked to explain why they decided not to use the tool and the most common reasons were as follows: (a) they did not think they needed it, (b) they did not find it to be helpful, (c) they preferred to use other resources that were available, and (d) they found it confusing or difficult to use. A few mentioned that it did not work for them, perhaps due to technical errors. A very small number of students mentioned that they dislike AI and/or did not believe it should be implemented in academic settings.

Some discrepancies were observed between students' actual AI tutor usage and their survey responses. Among students who reported not using the AI tutor at all, 71 out of 111 had actually used it at least once during the semester, with usage ranging from 0.45% to 43.44% of the homework problems (Table 3). Despite this, these students still reported no AI usage. Conversely, five students who indicated in the survey that they had used the AI tutor did not use it on any problems.

**Table 3: Summary statistics of AI tutor use among students who reported no use**

| Variable | N | Mean | Std. Dev | Min | Pctl. 25 | Pctl. 50 | Pctl. 75 | Max |
|---|---|---|---|---|---|---|---|---|
| Percent of problems with AI use | 71 | 3.31% | 7.46% | 0.45% | 0.90% | 1.36% | 2.71% | 43.44% |

## 3.2 Demographic and Academic Characteristics Associated with Use of an AI Tutor

To investigate which student characteristics were associated with use of an AI tutor, models were run using stepwise variable selection with AIC as the criterion on the following variables: cumulative GPA, gender, underrepresented minority (URM) status, first-generation status, and year in school.

*3.2.1 Characteristics Associated with Use of the AI Tutor.* A logistic regression model was run to examine if student characteristics were associated with use of the AI tutor. The final model included all characteristics except URM status and is shown in Table 4. Results indicated that both female ($p$ = 0.006) and first and second year students ($p$ = 0.026) are more likely than are male and third and fourth year students to use the AI tutor at least once during the semester. However, the overall model fit was poor and the individual effect sizes were small, suggesting that these characteristics are not meaningfully related to whether or not a student will adopt an AI tutor.

*3.2.2 Characteristics Associated with Frequency of AI Use.* A regression model was run to examine if student characteristics were associated with how frequently students utilized the AI tutor. The final model included only gender and is shown in Table 5. Results indicated that female students used the AI tutor for more problems than did male students ($p$ = 0.0429). However, the overall model fit was poor, suggesting that gender explains an insignificant portion of the variance in the frequency of AI use.

## 3.3 Relationship between Frequency of AI Tutor Use and Performance

In order to investigate the potential association between using the AI tutor and performance in the course, two linear regression models were fit, which included incoming GPAs and the total number of homework problems attempted as covariates. All 660 participants were considered for these models regardless of whether they used AI at any point in the semester or not, but 13 did not have incoming GPAs and were excluded from the models. The first model examined performance on assigned homework problems (Table 6). In this model, the percentage of homework problems during which students used AI at any point was a significant positive predictor of overall homework performance.

The second model examined overall exam performance (Table 7). In this model, the percentage of homework problems students attempted during which they used AI at any point was not a significant predictor of overall exam performance. Thus, we found evidence of an association between using the AI tutor for assigned homework problems and homework performance but did not find evidence of an association with exam performance.

## 3.4 Patterns of AI Tutor Use

*3.4.1 Problem-Level Clusters.* Analysis of the 44 problems across the three most AI assisted homework assignments resulted in the identification of seven distinct clusters. Seven was determined to be the optimal number of problem level clusters as this was the point at which the further increases in the number of clusters lead to a smaller increase in the mean silhouette. Beyond this point, increasing the number of clusters tended to obfuscate any discernible trends. The mean silhouette value for the seven-cluster solution was 0.92 indicating good clustering structure (Figure 2). Process maps were then used to visualize the most frequent 90% of traces for each of these seven problem-level clusters (Figure 3). The clear homogeneity visible in the traces of these clusters supports the validity of this clustering solution.

**Problem Cluster 1: Correct Initial Attempt without AI (47.1% of traces)** This cluster is comprised of traces with only a single correct attempt. For many problems, students were able to get the correct answer on their initial attempt, without any assistance from the AI tutor. There were 15,773 traces representing this cluster, representing almost half of the traces in the sample.

**Problem Clusters 2-4: Incorrect Attempts (32.6% of traces)** Problem Clusters 2-4 primarily represent instances where students were not able to get the correct answer on their initial attempt, but did not utilize the AI tutor for help. For problem Clusters 2 and 3, the traces include a correct submission after 1 or more incorrect attempts, indicating that students eventually got the correct answer. There were 7,046 (21.0%) and 2,483 (7.4%) traces representing Clusters 2 and 3 respectively. There were 1,370 traces (4.1%) in Cluster 4, which is primarily made up of traces with no correct submissions. Thus, Cluster 4 represents instances when a student used up their three allotted attempts or gave up without getting the correct answer.

**Problem Cluster 5: Chat after Incorrect Attempts (6.2% of traces )** This cluster is represented by cases where students submitted one or more incorrect attempts, and then asked the AI tutor for help.

**Problem Cluster 6: Helpful Initial Chat (8.9% of traces)** This cluster consists of cases where students asked the AI tutor for help before submitting any answer and then had a correct submission.

**Problem Cluster 7: Less helpful Initial Chat (5.2% of traces)** This cluster consists of cases where students asked the AI tutor for help before submitting any answer but then had at least one incorrect attempt. In most cases, they eventually got the correct answer.

*3.4.2 Sequence-Level Clusters.* After classifying all problem traces into one of the seven problem clusters described above, the next step was to determine whether specific patterns emerged related to how students engaged with these homework problems across multiple problems and assignments. Student-level clusters were generated

**Table 4: Final model for demographics associated with use of the AI tutor**

| Variable | B | OR | SE |
|---|---|---|---|
| Constant | 1.65* | | 0.76 |
| GPA | 0.39 | 1.484 | 0.22 |
| Gender (Male)[a] | -0.88** | 0.413 | 0.32 |
| First.Generation[b] | 0.64 | 1.891 | 0.43 |
| Academic.Year (3rd or 4th)[c] | -0.88* | 0.416 | 0.39 |
| Pseudo.R2 | .06 | | |
| N | 641 | | |

*Note.* [a]: Female = 0, Male = 1
[b]: Not First-Generation = 0, First Generation = 1
[c]: 1st or 2nd Year = 0, Not 1st or 2nd Year = 1
*p < .05, **p < .01

**Table 5: Final model for demographics associated with frequency of use of the AI tutor**

| Variable | B | $\beta$ | SE |
|---|---|---|---|
| Constant | 14.97*** | | 0.59 |
| Gender[a] | -2.24* | -0.08 | 1.10 |
| Adj.R2 | .01 | | |
| N | 596 | | |

*Note.* [a]: Female = 0, Male = 1*p < .05 ***p < .001

**Table 6: Regression model for frequency of AI tutor use on homework scores**

| Variable | B | $\beta$ | SE |
|---|---|---|---|
| Constant | -19.47*** | | 2.51 |
| GPA | 2.56*** | 0.2 | 0.26 |
| Total Homework Problems Attempted | 0.48*** | 0.79 | 0.01 |
| Percent of Problems with AI use | 0.03** | 0.05 | 0.01 |
| Adj.R2 | .77 | | |
| N | 647 | | |

*Note.* **p < .01 ***p < .001

**Table 7: Regression model for frequency of AI tutor use on exam scores**

| Variable | B | $\beta$ | SE |
|---|---|---|---|
| Constant | 21.34* | | 8.7 |
| GPA | 18.86*** | 0.66 | 0.9 |
| Total Homework Problems Attempted | -0.01 | -0.01 | 0.04 |
| Percent Homework Problems AI | -0.01 | 0.00 | 0.04 |
| Adj.R2 | .43 | | |
| N | 647 | | |

*Note.* *p < .05 ***p < .001

using K-medoids clustering. A local maximum of the average silhouette value was used to select five clusters. The problem-level clusters sequenced across the 44 problems can be seen in Figure 4. The mean silhouette value of .26 suggested a weak clustering effect. However, we did not expect students interactions across 44 homework problems to be strongly homogeneous, so we wanted to explore any potential utility in this clustering to understand

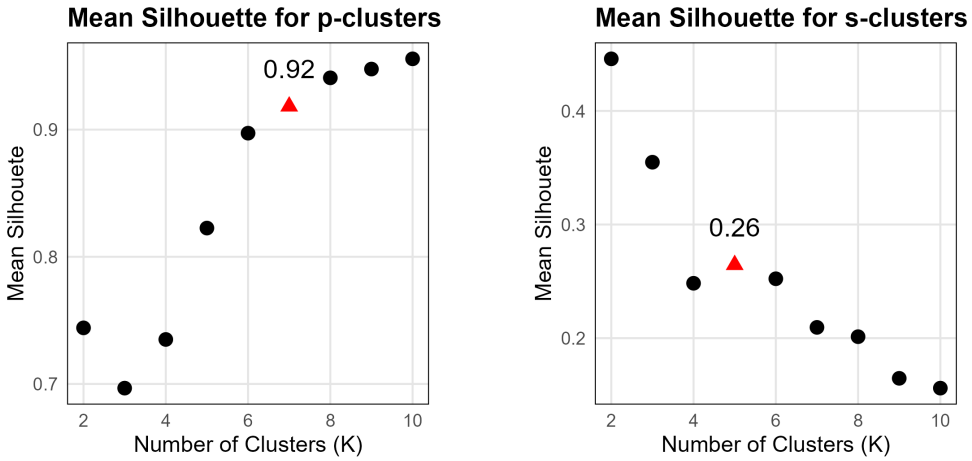


**Figure 2: Average silhouette values for problem-level (left) and student-level (right) clustering using 1 to 10 clusters. The selected number (K) of clusters is indicated by a red triangle, and the average silhouette of the selected K is displayed.**

student learning. The distribution of p-clusters within each of the five s-clusters can be seen in Table 8

**Sequence Cluster 1: *Low Performers* (N = 84)** - Students in this cluster had the lowest percentage of correct initial attempts (42%) and lowest percentage of problems that were correct within three attempts (88%) out of the five groups. This group had the second lowest AI use; they utilized the tutor on 5.6% of the problems they attempted in the clustered set and 5.1% across all the assigned homework problems.

**Sequence Cluster 2: *Non-chatters* (N = 200)** – Students in this cluster rarely used the AI tutor using it for only 3.8% of the problems they attempted in the clustered set and 3.9% across all the assigned homework problems. They tended to get their initial attempts correct (62%), typically without the use of the AI tutor; the most frequent problem cluster among this group was correct initial attempt (61%).

**Sequence Cluster 3: *Chatters* (N = 54)** – Overall, students in this cluster used the AI tutor extensively. The most frequent problem cluster among this group was helpful chat (36%), and they typically messaged the tutor prior to submitting any attempts. Overall, they utilized the tutor on 65% of the problems they attempted in the clustered set and 39% across all the assigned homework problems. On problems where they used the AI tutor, they also exchanged more messages than the other four groups, sending on average seven messages per problem, and they used the premade prompts less frequently than the other four clusters, typing their own messages 87% of the time.

**Sequence Cluster 4: *Mixed Approach* (N = 137)** - The most frequent problem cluster among this group was a correct initial attempt (37.4%). They utilized the AI tutor less frequently than Cluster 3, but more than the other four clusters. Overall, they utilized the tutor on 39% of the problems they attempted in the clustered set and 22% across all the assigned homework problems. They were somewhat more likely to utilize the AI prior to submitting an attempt than to ask the tutor for help after they had already submitted one or more incorrect attempts, but not to the same extent as Cluster 3. Despite their use of the AI tutor, this group had the second lowest percentage of problems that were correct within three attempts (93%) out of the five groups

**Sequence Cluster 5: *Selective Chatters* (N = 127)** – This group used the AI tutor for 14% of problems across all problems and 23% of problems in the clustered problem set. These students got the majority of their initial attempts correct (60%), and they typically did so without the help of the AI tutor. The most frequent problem cluster among this group was correct initial attempt (51%). They were more likely to use the AI tutor when they had already submitted one or more incorrect attempts than prior to submitting a problem attempt. They used the AI tutor in 22% of cases where they had already submitted two incorrect attempts.

*3.4.3 Differences in Performance Outcomes by S-Cluster Membership.* Exploratory ANOVA tests were run to determine if cluster membership was related to course outcomes. All results from the ANOVA tests are shown in Table 9. Results indicated significant differences in the s-clusters in terms of average homework scores, $F(4, 597) = 24.5$, FDR adj. $p < .001$. Post-hoc pairwise comparisons using Tukey HSD tests indicated that Cluster 1 (low performers) earned significantly lower homework scores than the other four groups. In addition, Cluster 4 (mixed approach) earned significantly lower homework scores than Clusters 2, 3, and 5.

Additionally, controlling for homework scores, results from ANOVA indicated significant differences in the s-clusters in terms of average exam scores, $F(4, 596) = 6.22$, FDR adj. $p < .001$. Post-hoc pairwise comparisons using Tukey HSD tests indicated that Cluster 5 earned higher exam scores than Clusters 1, 3, and 4.

## 4 DISCUSSION

Student overall use of the AI tutor was relatively low, though most students (93% of the participants) interacted with it at least once. The homework assignments, designed to promote learning, allowed multiple attempts to achieve maximum points, resulting in little

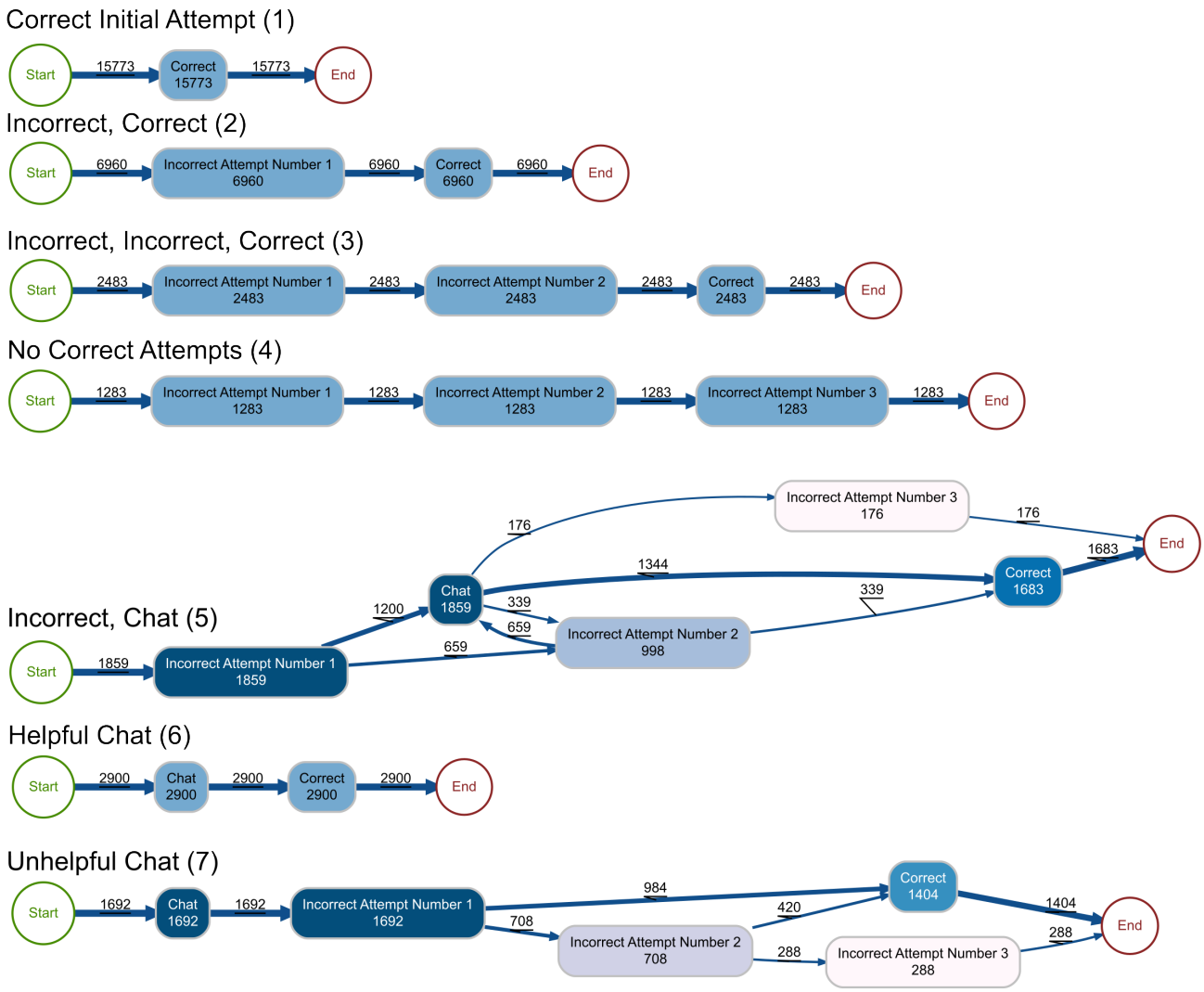


**Figure 3: Process maps showing the 90% most frequent traces for the problem-level clusters.**

**Table 8: Problem-cluster frequencies (%) for each of the five sequence clusters**

| S-Cluster | P-Cluster 1 | P-Cluster 2 | P-Cluster 3 | P-Cluster 4 | P-Cluster 5 | P-Cluster 6 | P-Cluster 7 |
|---|---|---|---|---|---|---|---|
| Cluster 1 | 40.9% | 29.0% | 14.0% | 10.6% | 3.2% | 1.0% | 1.3% |
| Cluster 2 | 60.8% | 23.8% | 8.2% | 3.3% | 2.0% | 0.9% | 0.9% |
| Cluster 3 | 24.4% | 8.2% | 1.7% | 0.8% | 11.2% | 36.4% | 17.5% |
| Cluster 4 | 37.4% | 16.4% | 4.9% | 2.3% | 11.7% | 17.0% | 10.4% |
| Cluster 5 | 50.9% | 19.7% | 5.0% | 1.5% | 8.2% | 9.0% | 5.6% |
| Overall | 47.1% | 21.0% | 7.4% | 4.1% | 6.2% | 8.9% | 5.2% |

variation in final homework scores, with an average of 93%. This could be one reason why most students used the AI tutor infrequently, as they felt they did not need it. However, when students did engage with the AI tutor, their interactions were more selective and involved extended dialogues rather than one-off queries, with an average chat length of 4.6 messages.

Students generally viewed the AI tutor as helpful, consistent with previous research [3]. While demographic factors like gender

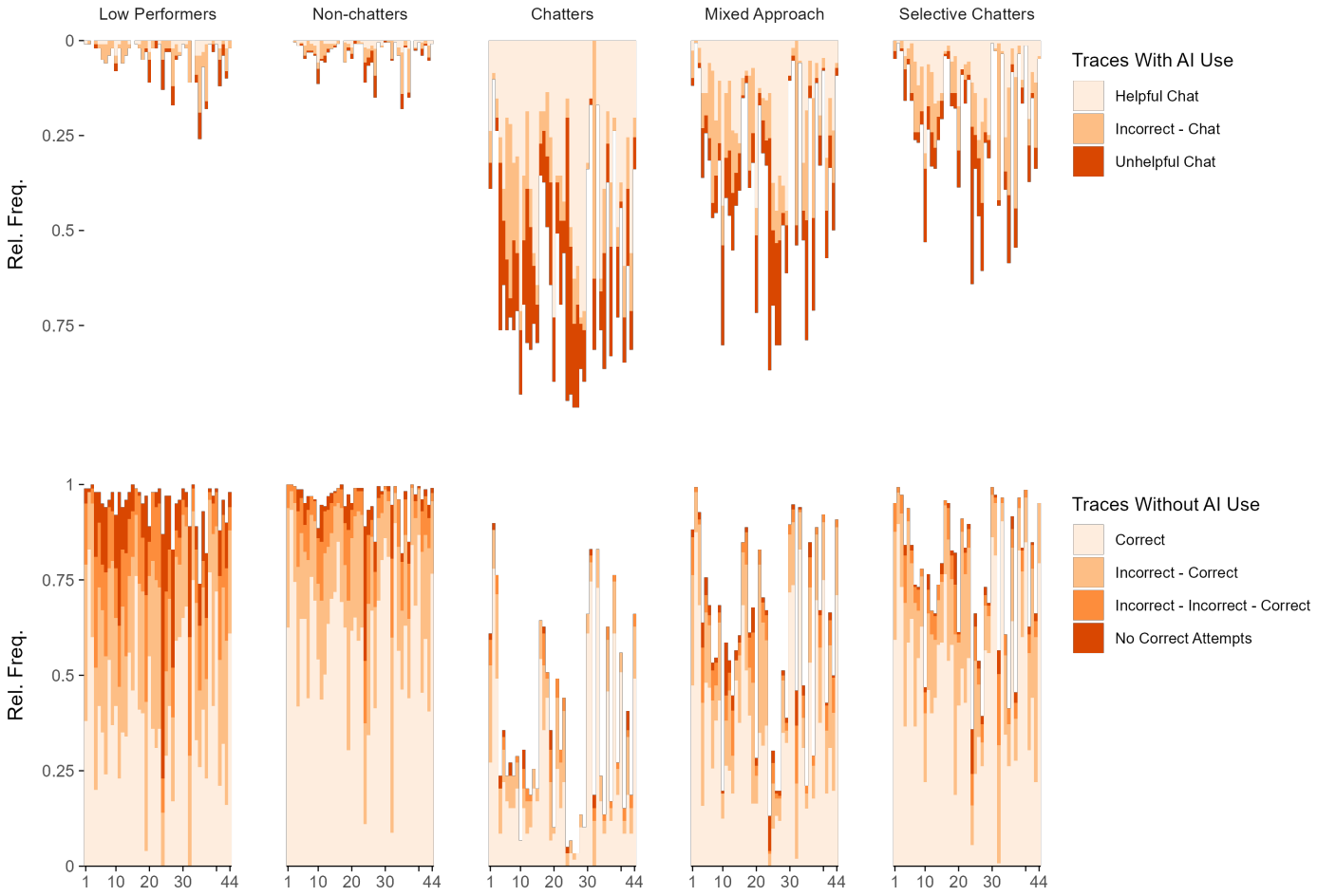


**Figure 4: Student level s-cluster sequence density plots separated by traces that used the AI tutor (top) and traces that did not (bottom). The sequences are ordered by the first event timestamp of the middle trace for each homework problem, and thus are roughly in chronological order.**

**Table 9: ANOVA results for performance outcomes by cluster**

| S-Cluster | N | GPA M(*SD*) | Homework Scores M(*SD*) | Exam Scores M(*SD*) |
|---|---|---|---|---|
| Cluster 1 | 84 | 3.16 (.61)$^{c}$ | 90.1 (6.0)$^{c}$ | 57.7 (13.8)$^{b}$ |
| Cluster 2 | 200 | 3.41 (.65)$^{b}$ | 95.4 (4.4)$^{a}$ | 67.3 (14.2)$^{ab}$ |
| Cluster 3 | 54 | 3.22 (.61)$^{bc}$ | 95.6 (3.4)$^{a}$ | 63.1 (11.6)$^{b}$ |
| Cluster 4 | 137 | 3.37 (.61)$^{bc}$ | 93.3 (7.0)$^{b}$ | 64.1 (14.6)$^{b}$ |
| Cluster 5 | 127 | 3.67 (.37)$^{a}$ | 96.3 (3.0)$^{a}$ | 72.0 (12.1)$^{a}$ |

*Note.* Different superscripts in the same column differ significantly ($p < .05$)

and academic year initially appeared to be associated with AI tutor use, the effect size was very small. We did not find evidence that these characteristics were meaningfully associated with usage. This suggests that other factors may play a larger role in student engagement with the AI tutor.

Although student use of the AI tutor was low on average (accounting for 14% of the homework problems), a positive association was observed between the number of homework problems AI assistance was used for and homework scores. Yet, we were not able to find evidence that indicated any kind of association between

frequency of AI use and exam performance. One might intuitively assume that students who used the AI tutor merely to obtain correct answers could increase homework scores without truly understanding the underlying concepts and subsequently perform worse on exams. Conversely, if students used the AI tutor to deepen their understanding of important concepts, this learning could potentially transfer to higher-stakes assessments, such as exams. Yet, we did not find a clear positive or negative association between frequency of AI tutor use and overall exam performance and this aligns with previous research [15].

When analyzing student engagement with homework, seven distinct problem patterns emerged. Four involved no interaction with the AI tutor, even after failed attempts. Among the three patterns with AI chats, interactions varied: some students used the AI tutor before attempting the problem, while others sought its help after an initial failure. Use of the AI tutor did not always lead to success; asking for help prior to attempting the problem resulted in a successful initial attempt just over half the time. Despite varied interaction patterns, five distinct student clusters emerged based on dominant behavior patterns.

The results suggest that AI tutor interaction frequency was not directly linked to exam performance outcomes. Examining the student-level clusters offers some potential explanations. The highest performing students used the AI tutor selectively, primarily using the chat after one or more unsuccessful attempts. These higher achieving students were more likely to seek help after making an initial attempt on their own, while lower achieving students were more likely to ask for help before attempting the problem. Surprisingly, some students, despite multiple failed attempts, rarely used the tutor, and this group exhibited the lowest learning outcomes. With the proper guidance and encouragement, these students could potentially benefit from the AI tutor.

Still, other students avoided using the AI tutor, but were able to solve problems independently and performed well on exams. It remains unclear, however, whether these students utilized other resources in place of the AI tutor. Additionally, some students heavily engaged with the AI tutor through frequent, extended dialogues, often seeking help before attempting problems. However, their exam performance was not higher than that of students who used the AI tutor less frequently but more selectively. This suggests that success in subsequent submissions may depend on how effectively students utilize the AI tutor's responses in conjunction with their prior knowledge.

Notably, behavior data is crucial to understanding the full scope of how students engage with the AI tutor. This study identified slight discrepancies between student behavior data and self-reported responses regarding AI use. While understanding students' perceptions and attitudes is important, self-reported data can be unreliable [5] and should be corroborated with actual behavior data for a more accurate understanding of AI tutor usage.

## 5 LIMITATIONS AND CONCLUSION

One key limitation of this study is uncertainty regarding the quality of the AI's assistance. Previous research has shown that students often prefer AI tutors over teaching assistants or instructors due to convenience, despite acknowledging that AI-generated help may be of lower quality [12]. However, as AI technology advances, these perceptions could change.

Another limitation is the lack of examination of other factors influencing AI usage, such as students' prior experiences or attitudes toward AI. Future studies should explore these aspects to provide a more comprehensive understanding of AI adoption and engagement behaviors in educational contexts.

Despite these limitations, this study provides valuable insights into the diverse ways students interact with an AI tutor, linking their interaction pattern to their prior learning and performance outcomes. Although the AI tutor was available without restrictions, most students opted to complete each homework problem independently, using the AI selectively if at all. High-achieving students tended to use the AI tutor selectively, while the lowest performing students often did not seek help from the AI tutor when help was needed. A small group of students, however, engaged extensively with the AI tutor.

Overall, this study offers important insights for exploring AI integration in educational settings and suggests that students could benefit from guidance on how to use the AI tutor effectively. When used strategically, the AI tutor could be a valuable tool for enhancing student learning.

## Acknowledgments

We acknowledge Macmillan Learning for providing the student interaction data utilized in this study. Collected through the Achieve homework platform used in the course, these data enabled our analysis of student engagement with the AI tutor and its impact on performance outcomes.

## References

[1] Hüseyin Ateş. 2024. Integrating augmented reality into intelligent tutoring systems to enhance science education outcomes. *Education and Information Technologies* (2024), 1–36.

[2] Beatriz Borges, Negar Foroutan, Deniz Bayazit, Anna Sotnikova, Syrielle Montariol, Tanya Nazaretzky, Mohammadreza Banaei, Alireza Sakhaeirad, Philippe Servant, Seyed Parsa Neshaei, et al. 2024. Could ChatGPT get an Engineering Degree? Evaluating Higher Education Vulnerability to AI Assistants. *arXiv preprint arXiv:2408.11841* (2024).

[3] Cecilia Ka Yuk Chan and Wenjie Hu. 2023. Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education* 20, 1 (2023), 43.

[4] Marcia Devlin and Kathleen Gray. 2007. In their own words: A qualitative study of the reasons Australian university students plagiarize. *High Education Research & Development* 26, 2 (2007), 181–198.

[5] Charles Dinerstein. 2019. Measuring the Reliability of Self-Reported Behavior. *American Council on Science and Health* (2019). https://www.acsh.org/news/2019/09/25/measuring-reliability-self-reported-behavior-14301 Accessed: 2023-09-23.

[6] Lu Ding, Tong Li, Shiyan Jiang, and Albert Gapud. 2023. Students' perceptions of using ChatGPT in a physics class as a virtual tutor. *International Journal of Educational Technology in Higher Education* 20, 1 (2023), 63.

[7] Tom Farrelly and Nick Baker. 2023. Generative artificial intelligence: Implications and considerations for higher education practice. *Education Sciences* 13, 11 (2023), 1109.

[8] Eduard Frankford, Clemens Sauerwein, Patrick Bassner, Stephan Krusche, and Ruth Breu. 2024. AI-Tutoring in Software Engineering Education. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering Education and Training*. 309–319.

[9] Alexis Gabadinho, Gilbert Ritschard, Nicolas S Müller, and Matthias Studer. 2011. Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software* 40, 4 (2011), 1–37. https://doi.org/10.18637/jss.v040.i04

[10] Kevin Gold and Shuang Geng. 2024. Tracking the Evolution of Student Interactions with an LLM-powered Tutor. In *Proceedings of the International Conference on Learning Analytics and Knowledge (LAK) '24*. ACM Press, New York.

[11] Shikoh Hirabayashi, Rishab Jain, Nikola Jurković, and Gabriel Wu. 2024. Harvard Undergraduate Survey on Generative AI. *arXiv preprint arXiv:2406.00833* (2024).

[12] Irene Hou, Sophia Mettille, Owen Man, Zhuo Li, Cynthia Zastudil, and Stephen MacNeil. 2024. The Effects of Generative AI on Computing Students' Help-Seeking Preferences. In *Proceedings of the 26th Australasian Computing Education Conference*. 39–48.

[13] Heather Johnston, Rebecca F Wells, Elizabeth M Shanks, Timothy Boey, and Bryony N Parsons. 2024. Student perspectives on the use of generative artificial intelligence technologies in higher education. *International Journal for Educational Integrity* 20, 1 (2024), 2.

[14] Kenneth R. Koedinger and Vincent Aleven. 2016. An integrated vision for intelligent tutoring systems: Advancing the science of learning and instruction by merging cognitive modeling with machine learning. *International Journal of Artificial Intelligence in Education* 26, 1 (2016), 238–256. https://doi.org/10.1007/s40593-015-0079-4

[15] Tomaž Kosar, Dragana Ostojić, Yu David Liu, and Marjan Mernik. 2024. Computer Science Education in ChatGPT Era: Experiences from an Experiment in a Programming Course for Novice Programmers. *Mathematics* 12, 5 (2024), 629.

[16] Kunyanuth Kularbphettong, Pubet Kedsiribut, and Pattarapan Roonrakwit. 2015. Developing an adaptive web-based intelligent tutoring system using mastery learning technique. *Procedia-Social and Behavioral Sciences* 191 (2015), 686–691.

[17] James A Kulik and John D Fletcher. 2016. Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of educational research* 86, 1 (2016), 42–78.

[18] Mengqi Liu and Faten M'hiri. 2024. 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*. 743–749.

[19] Iris Ma, Alberto Krone Martins, and Cristina Videira Lopes. 2024. Integrating AI Tutors in a Programming Course. *arXiv preprint arXiv:2407.15718* (2024).

[20] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. 2023. *cluster: Cluster Analysis Basics and Extensions.* https://CRAN.R-project.org/package=cluster R package version 2.1.6 — For new features, see the 'NEWS' and the 'Changelog' file in the package source).

[21] Ha Nguyen, Nate Stott, and Vicki Allan. 2024. Comparing Feedback from Large Language Models and Instructors: Teaching Computer Science at Scale. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*. 335–339.

[22] OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/research/ Accessed: 2124-09-11.

[23] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[24] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[25] Odin Monrad Schei, Anja Møgelvang, and Kristine Ludvigsen. 2024. Perceptions and Use of AI Chatbots among Students in Higher Education: A Scoping Review of Empirical Studies. *Education Sciences* 14, 8 (2024), 922.

[26] Andreas Scholl, Daniel Schiffner, and Natalie Kiesler. 2024. Analyzing Chat Protocols of Novice Programmers Solving Introductory Programming Tasks with ChatGPT. *arXiv preprint arXiv:2405.19132* (2024).

[27] James S Sharpe, Ryan E Dougherty, and Sarah J Smith. 2024. Can ChatGPT Pass a CS1 Python Course? *Journal of Computing Sciences in Colleges* 39, 8 (2024), 128–142.

[28] Adele Smolansky, Andrew Cram, Corina Raduescu, Sandris Zeivots, Elaine Huber, and Rene F Kizilcec. 2023. Educator and student perspectives on the impact of generative AI on assessments in higher education. In *Proceedings of the tenth ACM conference on Learning@ Scale*. 378–382.

[29] Marek Urban, Filip Děchtěrenko, Jiří Lukavskỳ, Veronika Hrabalová, Filip Svacha, Cyril Brom, and Kamila Urban. 2024. ChatGPT improves creative problem-solving performance in university students: An experimental study. *Computers & Education* 215 (2024), 105031.

[30] Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist* 46, 4 (2011), 197–221.

[31] Candace Walkington and Matthew L Bernacki. 2019. Personalizing algebra to students' individual interests in an intelligent tutoring system: Moderators of impact. *International Journal of Artificial Intelligence in Education* 29 (2019), 58–88.

[32] Karen D Wang, Eric Burkholder, Carl Wieman, Shima Salehi, and Nick Haber. 2024. Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving. In *Frontiers in Education*, Vol. 8. Frontiers Media SA, 1330486.

[33] Beverly Park Woolf. 2010. *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning.* Morgan Kaufmann.

[34] Tom Zhang, Michelle Taub, and Zhongzhou Chen. 2022. A multi-level trace clustering analysis scheme for measuring students' self-regulated learning behavior in a mastery-based online learning environment. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. 197–207.