The Effects of Chatbot Placement, Personification, and Functionality on Student Outcomes in a Global CS1 Course

Sierra Wang
sierraw@cs.stanford.edu
Stanford University
CA, USA

Thomas Jefferson
tjj@stanford.edu
Stanford University
CA, USA

Chris Piech
piech@cs.stanford.edu
Stanford University
CA, USA

John C Mitchell
john.mitchell@stanford.edu
Stanford University
CA, USA

ABSTRACT

We conducted a large-scale randomized controlled trial (RCT) in a massive, open-access, online CS1 course to examine how a chatbot's (1) placement within the course interface, (2) degree of personification, and (3) technical functionality impact student outcomes. When the chatbot was placed in the course's Integrated Development Environment instead of the lessons interface, over five times the proportion of students sent a message. Additionally, these students sent four times as many messages and asked about the homework four times as often. Subtle design choices – such as greeting students with "Hello" – nearly doubled the percentage of students who engaged with the tool. Despite these significant differences in chatbot usage, we found no meaningful impact on student performance in the course. The study included 6,515 students from 149 countries, divided into 10 experiment groups – each receiving a distinct chatbot implementation – and one control group. Our results show that chatbot design dramatically influences how students engage with the tool, providing valuable insights for further development of AI-based pedagogical tools.

CCS CONCEPTS

• **Social and professional topics** → *CS1*; • **Computing methodologies** → *Natural language generation*; • **Applied computing** → **Computer-assisted instruction**.

KEYWORDS

Randomized Controlled Trial, Chatbot, AI, GPT, CS1, Global

ACM Reference Format:
Sierra Wang, Thomas Jefferson, Chris Piech, and John C Mitchell. 2025. The Effects of Chatbot Placement, Personification, and Functionality on Student Outcomes in a Global CS1 Course. In *Proceedings of the Twelfth ACM Conference on Learning @ Scale (L@S '25), July 21–23, 2025, Palermo, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3698205.3729557

L@S '25, July 21–23, 2025, Palermo, Italy

ACM ISBN 979-8-4007-1291-3/2025/07
https://doi.org/10.1145/3698205.3729557

1 INTRODUCTION

Generative AI offers unprecedented opportunities to support students in their learning, but may also distract or demotivate them from meaningful educational experiences. For example, a recent study [34] suggests that course-based access to ChatGPT may both reduce student engagement *and* increase exam scores. These complex effects necessitate thoughtful experimentation with new AI tools in order to carefully examine the nuanced impacts of different approaches. In our effort to promote access to meaningful and effective computer science education worldwide, we conducted a large-scale randomized controlled trial (RCT) exploring the integration of generative AI-based chatbots in a global CS1 course. We aimed to understand how novice programming students engage with chatbots, how these interactions impact their outcomes, and how these tools can be improved to better support learning. Our study involved 6,515 students from 149 countries, divided into experiment groups—each receiving a distinct chatbot implementation—and a control group that did not receive a chatbot.

Based on the growing literature of generative AI chatbots in learning environments (see Section 2), we posed a series of preregistered hypotheses on how different features of a chatbot could impact learning. To test these hypotheses, we developed multiple chatbot implementations, strategically varying their placement in the course website, level of personification, and technical functionality. Chatbots were placed alternatively in the lessons interface, where students watch lecture videos, or in the course's Integrated Development Environment (IDE), where they solve programming problems. Some students were able to chat through free-form messages, while others were simply offered buttons labeled "Explain it differently" and "Why does it matter?" Some were greeted with "Hello! I'm Chat, the AI chatbot"; others were told it was a "space to ask questions." For some students, the chatbot functionality was engineered using Retrieval-Augmented Generation (RAG) techniques to direct them toward relevant course materials; for others, it broadcast their messages to a community of learners to foster collaborative learning experiences. Some students did not have access to an AI based chat feature at all.

We examine how these design choices influence student engagement and learning outcomes by systematically comparing the impact of these varied implementations. We find that placing the chatbot in the IDE, as opposed to the lessons interface, dramatically impacts how students use the tool – five times the proportion of

students sent a message and those that did sent four times as many messages. Additionally, these messages were about the homework four times as often. Subtler differences in presentation – such as the chatbot greeting students with "Hello" – nearly doubled engagement with the tool. Despite these differences in usage, we find a surprising null result: inclusion of an AI-based chatbot, regardless of the implementation, did *not* improve student learning outcomes. Although one might expect the technical functionality of a chatbot (e.g., course-based RAG) to significantly influence its impact on students, our study dramatically reveals how presentational choices can overshadow these technical factors. As generative AI continues to shape learning experiences, our results demonstrate that thoughtful integration – not just access to technology – will be key to maximizing its educational benefits.

Main Contributions:

(1) **Large-Scale RCT on Chatbot Integration:** We designed, preregistered, and conducted a large-scale RCT involving 6,515 students from 149 countries in an online CS1 course. Our study examines four key dimensions of chatbot integration: presence, placement in the course interface, personification of AI, and technical functionality.
(2) **Impact of Placement and Personification:** Our findings reveal that chatbot placement and personification significantly influence student engagement with the tool, often in ways that contradict our expectations. These insights have critical implications for instructional design.
(3) **Surprising Null Result:** Although the chatbots' designs dramatically impacted how students used them, no chatbot implementation significantly outperformed the control group in terms of course completion. In fact, the control group had slightly more course engagement.
(4) **Demographics Analysis:** We analyzed chatbot impact across student backgrounds and identified a gender bias in the impact of chatbot placement. Women had higher assignment completion rates with chatbots integrated into lessons, whereas men had higher completion with IDE placement.

Organization. We review related work in Section 2, describe our methodology in Section 3, results in Section 4, possible interpretations and demographics analysis in Section 5, limitations in Section 6, and conclude in Section 7.

2 RELATED WORK

The increasing power of large language models (LLMs) is both a source of opportunity and a topic of concern in education [13, 15, 16, 18, 32]. In particular, AI-based chat systems are being widely integrated into educational platforms [3–6], with hopes of greatly increasing access to personalized education [28, 49].

There is extensive prior work examining the impact of conversational agents and personal tutors in education based on earlier AI methods [12, 25, 42, 43, 48, 52]. More recently, several works have explored the educational implications of OpenAI's ChatGPT [7, 17, 22, 24, 36, 44, 46], while others have implemented their own custom chat systems [8, 10, 19, 23, 26, 29, 30, 45, 47, 53]. Through surveys, engagement metrics, and other methods, analyses have shown mixed results regarding the impact of these tools [9, 11, 14, 21, 33, 39, 41]. For example, Liu et al. deployed a chatbot in both a university setting and a massive online class, and found overall satisfaction and increasing adoption of the tool among students [27]. In a different study, Nie et al. examined the impact of a course based ChatGPT interface through a large scale RCT and found that simply offering an AI interface to students can reduce engagement [34]. These nuanced findings demonstrate the subtle and complex impacts of AI-based tools, highlighting the need for further experimentation and analysis in this area.

Although many have implemented AI chat systems in educational settings, the literature lacks a comprehensive comparative analysis of the placement, functionality, and personification of chat systems. This lack of clarity around which chatbot features enhance learning and which detract from it motivates the present study.

3 METHODOLOGY

We conducted a large scale randomized controlled trial (RCT) to test the following hypotheses:

(1) **H1**: Chatbots will increase lesson and assignment completion rates.
(2) **H2**: If the chatbot is placed within the lessons, then students will show higher lesson and assignment completion rates than if the chatbot is placed in the IDE.
(3) **H3**: Personification of AI will decrease lesson and assignment completion rates.

We preregistered our hypotheses, experiment, and analysis plan with Open Science Framework. Our complete preregistration is at https://doi.org/10.17605/OSF.IO/QXAZ6. We describe the course context (Section 3.1), chatbot variations (Section 3.2), evaluation metrics (Section 3.3), and analysis plan (Section 3.4) below.

3.1 Course Context and Participants

We ran our study in Code in Place 2024, a free, open-access, online, introductory computer science course that teaches the basics of Python to students around the world [2, 31, 38]. The course includes three primary components: lessons, assignments, and weekly discussion sections. In the lessons, students learn about new concepts by watching lecture videos and working through guided exercises. In the assignments, students strengthen their understanding of these concepts by writing code in the Integrated Development Environment (IDE) that is embedded in the course website. In the sections, students meet in groups of about 10 students with a section leader to review the course concepts. The sections occur once a week and are the only synchronous component of the course.

All students who enrolled in the course were initially included in the experiment. We used cluster randomization to assign students to experiment groups, where each discussion section was randomly assigned to an experiment group. In other words, all of the students in a given discussion section were assigned to the same experiment group. We did this to prevent students from learning about the different experiment arms and maintain the integrity of the RCT. If a student changed discussion sections during the course, they are excluded from all results presented in this paper. To ensure that this exclusion did not influence our results, we repeated our analysis including these students and provide these results at https://github.com/sierrawang/cip-chatbots.git.

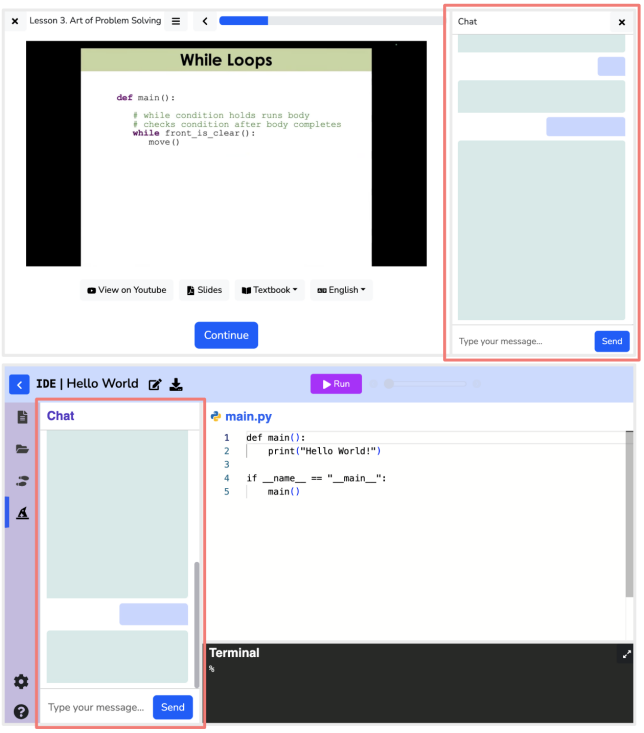

Figure 1: The top image shows the chatbot integrated in the lessons interface (alongside a lecture video), and the bottom image shows the chatbot integrated in the course IDE.

3.2 Chatbot Variations

We developed ten distinct chatbot implementations that varied in (1) placement within the course interface, (2) personification of AI, and (3) technical functionality. This allowed us to systematically compare the impact of each dimension. Here, we describe each dimension in detail:

(1) Placement: Lessons vs. IDE

We hypothesized that placing the chatbot either within the lessons interface or the IDE would significantly influence how students use the tool and, consequently, their learning outcomes. To understand the impact of chatbot placement on students, we integrated the chatbot interface within the IDE for two groups, and in the lessons interface for all other experiment groups. Figure 1 displays the chat interface in each location.

(2) Personification: Agent vs. Tool

Prior research on AI based educational tools has shown mixed results; while some tools have improved student outcomes [50, 51], others have led to student disengagement [34]. We hypothesized that the personification of AI influences how students perceive an AI based tool and may be causing these contrasting effects. To investigate this, we evaluate two levels of personification for the chatbot: "Agent" and "Tool". An Agent chat greets students with a message like, "Hello! I'm Chat, the AI chatbot for this course." If a student asks who they are chatting with, the Agent chat responds that it is "an AI chatbot." In contrast, the Tool chatbot introduces the interface by saying, "This is your space to ask questions" and describes itself as "an educational tool created using GPT-4 technology." We compare the impact of personification by giving half of the experiment groups an Agent chat and the others a Tool chat.

(3) Functionality: Basic, RAG, Buttons, or Community

The implementation of an AI based chatbot system defines how the chatbot responds, and consequently, influences how students use it. To understand the benefits of different approaches, we developed and assessed the following distinct implementations:

- **Basic**: The chatbot is implemented through basic prompting of the LLM to help the student understand the programming concepts involved in their problem. This implementation highlights the language model's default capabilities.
- **RAG**: The chatbot is implemented using Retrieval Augmented Generation (RAG) techniques to guide the student towards course materials. When a student asks a question, the system identifies the programming concepts involved, retrieves the most relevant and timely course materials, and uses them to craft a response. Through this implementation, we investigate whether it is beneficial for the LLM to use related course context when responding to a student, and whether it is advantageous to direct students towards the course materials.
- **Buttons**: This implementation allows students to interact with the chatbot through two buttons: "Explain it differently" and "Why does it matter?" We hypothesized that free form messages might not be the optimal format for students to seek help. With this implementation, we explore whether simple button presses could be a more effective interface.
- **Community**: Unlike the traditional 1-1 chatbot interaction, this implementation enables responses from both AI and student peers. When a student sends a message, the system sends it to other students who are studying the same material at the same time, in addition to the LLM. This setup aims to foster synchronous collaborations within an asynchronous learning environment. Collaborative learning can help students strengthen their understanding of concepts and feel a sense of community, which is particularly important in online contexts [20]. Through this implementation, we explore the possibility and efficacy of collaborative learning via AI based chatbots.

We examine these dimensions with ten experiment groups, each given a chatbot with a specific combination of placement, personification, and functionality. Two groups were assigned a chatbot with Basic functionality in the IDE, one with each level of personification (Agent and Tool). Eight groups received their chatbot in the lessons interface — two for each type of functionality (Basic, RAG, Community, and Buttons), with each level of personification. The control group, which did not receive a chatbot, helps us compare the impact of having any chatbot to not having one.

We implemented all of the chatbots using OpenAI's GPT-4 as the language model [35], and GPT-4-turbo specifically for the RAG implementation. Each prompt to the language model included context about the student's current task to ensure relevant and accurate responses. The code for each chatbot implementation is publicly available at https://github.com/sierrawang/cip-chatbots.git.

3.3 Evaluation Metrics

We utilize the following evaluation metrics to examine the impact of the different chatbot implementations on student behavior. We compute each metric at the individual student level, and to compare metrics across groups, we compute the mean for each group. We use bootstrapping to assess the statistical significance of the difference between group means [37]. Unless otherwise noted, we compute the mean across all members of the group.

Chat Usage Metrics. These metrics measure the extent of student engagement with the chatbot and provide insights into how students use them. *Sent Message* is a binary indicator for whether the student sent at least one message (1 if the student sent at least one message, 0 otherwise). *Message Count* is the total number of messages sent by the student. When comparing groups, the group mean is computed only for students who sent at least one message.

Course Completion Metrics. These metrics represent the completion rates of the primary course components, highlighting how the chatbots influence course engagement and outcomes. *Assignment Completion* is the percent of assignments that the student completed. *Lesson Completion* is the percent of lessons that the student completed. *Section Attendance* is the number of discussion sections that the student attended.

Forum Activity Metrics. These metrics assess how the chatbots impact student engagement with the course discussion forum, a separate resource where students can collaborate and receive help from the course staff. *Made Post* is a binary indicator for whether the student made at least one post on the discussion forum (1 if the student made at least one post, 0 otherwise). *Post Count* is the total number of posts made by the student. The group mean is computed only for students who made at least one post.

Exam Performance Metrics. These metrics estimate the impact of the chatbots on student learning as their performance on the exam. It is important to note that the exam was optional and graded by GPT-4 [1]. *Took Exam* is a binary indicator for whether the student took the exam (1 if the student took the exam, 0 otherwise). *Exam Score* is the exam score of the student. In an effort to address the missing exam scores, we perform two estimates when computing a group's results: (1) *Observed Score* is the mean exam score of students who submitted the exam. (2) *Adjusted Score* is the mean exam score with missing scores treated as zero. This is a lower bound for the group mean.

Confounding factors. The following demographic characteristics and chatbot variations could influence student behavior. When comparing groups, we report the group means for the relevant factors and the p-value for the observed difference. *Female* is a binary indicator for whether the student identifies as female. *Age* is the student's age. *In USA* is a binary indicator for whether the student lives in the United States of America. *Agent* is a binary indicator for whether the student received a chatbot personified as an Agent. *IDE* is a binary indicator for whether the student received a chatbot placed in the IDE. *Community* is a binary indicator for whether the student received a chatbot with Community functionality. *RAG* is a binary indicator for whether the student received a chatbot with RAG functionality. *Buttons* is a binary indicator for whether the student received a chatbot with Buttons functionality. *Control* is a binary indicator for whether the student was in the control group.

3.4 Preregistered Hypotheses

We focus on three hypotheses that explore the impact of chatbot presence, placement, and personification on student outcomes. By systematically testing these hypotheses, we aim to identify chatbot features that enhance — or hinder — student success in introductory computer science. We outline our preregistered hypotheses along with the methods used to test them below:

H1: Chatbots will increase lesson and assignment completion rates. To evaluate this hypothesis, we compare students who received a chatbot (Chatbot group) to those who did not receive a chatbot (No-Chatbot group). The **Chatbot group** consists of the two experiment groups that received a chatbot with Basic functionality in the lessons interface. The **No-Chatbot group** consists solely of the control group, which did not receive a chatbot. We test two null hypotheses for H1: (1) The Chatbot group does not have higher lesson completion rates than the No-Chatbot group. (2) The Chatbot group does not have higher assignment completion rates than the No-Chatbot group. H1 is supported if we reject both null hypotheses.

H2: If the chatbot is placed within the lessons, then students will show higher lesson and assignment completion rates than if the chatbot is placed in the IDE. To evaluate this hypothesis, we compare students who received a chatbot in the lessons interface (Lessons group) to those who received a chatbot in the IDE (IDE group). The **Lessons group** consists of the two experiment groups who received a chatbot with Basic functionality integrated into the lessons interface. The **IDE group** consists of the two experiment groups who received a chatbot with Basic functionality in the IDE. We test two null hypotheses for H2: (1) The Lessons group does not have higher lesson completion rates than the IDE group. (2) The Lessons group does not have higher assignment completion rates than the IDE group. H2 is supported if we reject both null hypotheses.

H3: Personification of AI will decrease lesson and assignment completion rates. To evaluate this hypothesis, we compare students who received a chatbot personified as an Agent (Agent group) to those who received a chatbot personified as a Tool (Tool group). The **Agent group** consists of the five experiment groups who received a chatbot personified as an Agent. The **Tool group** consists of the five experiment groups who received a chatbot personified as an Tool. We test two null hypotheses for H3: (1) The Tool group does not have higher lesson completion rates than the IDE group. (2) The Tool group does not have higher assignment completion rates than the IDE group. H3 is supported if we reject both null hypotheses.

Metric	Chatbot	No Chatbot	*p*-value
Chat Usage			
Sent Message	5.6%	N/A	N/A
Message Count	5.7	N/A	N/A
Course Completion			
Assn. Completion	60.8%	**61.4%**	0.763
Lssn. Completion	67.8%	**67.9%**	0.950
Sect. Attendance	69.6%	**70.6%**	0.535
Forum Activity			
Made Post	45.3%	**46.0%**	0.770
Post Count	**2.7**	2.6	0.548
Exam Performance			
Took Exam	23.0%	**24.1%**	0.587
Observed Score	77.5%	**80.9%**	0.121
Adjusted Score	17.8%	**19.5%**	0.321
Confounds			
Female	**50.1%**	45.0%	0.037
Age	32.4	**34.6**	0.000
In USA	29.5%	**34.4%**	0.032

Table 1: Compares students with a chatbot to those without a chatbot. For each comparison, the larger value is bold, and the p-value reflects the statistical significance of the differences.

Metric	Lessons	IDE	*p*-value
Chat Usage			
Sent Message	5.6%	**31.2%**	0.000
Message Count	5.7	**24.8**	0.003
Course Completion			
Assn. Completion	**60.8%**	59.0%	0.299
Lssn. Completion	**67.8%**	65.5%	0.103
Sect. Attendance	**69.6%**	68.3%	0.364
Forum Activity			
Made Post	**45.3%**	42.4%	0.156
Post Count	**2.7**	2.7	0.837
Exam Performance			
Took Exam	**23.0%**	21.0%	0.259
Observed Score	**77.5%**	77.2%	0.896
Adjusted Score	**17.8%**	16.2%	0.270
Confounds			
Female	**50.1%**	47.3%	0.179
Age	**32.4**	29.7	0.000
In USA	29.5%	**31.2%**	0.372
Agent	**52.1%**	46.8%	0.011

Table 2: Compares students with a chatbot in the lessons interface to those with a chatbot in the IDE. For each comparison, the larger value is bold, and the p-value reflects the statistical significance of the differences.

4 RESULTS

4.1 Chatbot or No Chatbot?

H1 — Chatbots will increase lesson and assignment completion rates — is not supported. Contrary to our hypothesis, the presence of a chatbot did not lead to improvements in course completion. Instead, the No-Chatbot group had slightly higher average assignment completion (61.4% vs. 60.8%; $p = 0.763$), lesson completion (67.9% vs. 67.8%; $p = 0.950$), and section attendance (70.6% vs. 69.6%; $p = 0.535$) rates compared to the Chatbot group. We also find that a higher proportion of the No-Chatbot group posted on the forum (46.0% vs. 45.3%; $p = 0.770$) and took the exam (24.1% vs. 23.0%; $p = 0.587$).

While none of these differences are statistically significant, the consistent direction of the results indicates that the chatbot did not increase student engagement by these metrics. Table 1 summarizes the observed metrics and statistical comparisons.

Notably, there are significant demographic differences between the Chatbot and No-Chatbot groups in gender, age, and U.S. residency as a result of our experiment group assignment protocol (see Section 3.1). To ensure these differences did not skew our results, we conducted additional regression analyses, detailed in Section 4.4, which confirm that there is no significant correlation between chatbot presence and any outcome metric. Altogether, we find no evidence to support H1.

Chatbots did not significantly improve outcomes. To further assess the effectiveness of AI-based chatbots, we compared the control group with each experiment group individually. When we compare the control group with the best performing experiment group for a metric, we find no significant difference, indicating that none of the chatbot designs offered a meaningful benefit for any measured outcome. Notably, the percent of students who used the chatbot was below 10% in the best-performing experiment group for each metric, suggesting that the tool may have had a limited role in influencing outcomes.

4.2 Chatbot Placement: Lessons or IDE?

H2 — If the chatbot is placed within the lessons, then students will show higher lesson and assignment completion rates than if the chatbot is placed in the IDE – is not supported. Placing the chatbot in the lessons interface yielded slightly higher assignment completion (60.8% vs. 59.0%; $p = 0.299$), lesson completion (67.8% vs. 65.5%; $p = 0.103$), and section attendance (69.6% vs. 68.3%; $p = 0.364$). Furthermore, the Lessons Group had higher rates of forum activity and exam performance, though these differences were not significant. Although these differences align with our hypothesis, they are not statistically significant, and therefore, this is insufficient evidence to support our hypothesis. Table 2 summarizes the observed metrics and statistical comparisons.

We note that the Lessons group had a significantly higher average age and proportion of students with an Agent-like chatbot. To evaluate how these imbalances impact our results, we conducted the regression analyses detailed in Section 4.4, which confirms that placing the chatbot in the IDE does not significantly correlate with course completion rates.

Metric	Agent	Tool	p-value
Chat Usage			
Sent Message	**12.2%**	7.7%	0.000
Message Count	**17.6**	12.6	0.146
Course Completion			
Assn. Completion	**60.4%**	58.8%	0.126
Lssn. Completion	**67.4%**	65.7%	0.045
Sect. Attendance	**68.9%**	68.3%	0.500
Forum Activity			
Made Post	**44.8%**	44.0%	0.583
Post Count	**2.8**	2.7	0.424
Exam Performance			
Took Exam	**22.1%**	21.4%	0.543
Observed Score	78.2%	**80.2%**	0.109
Adjusted Score	**17.2%**	17.2%	0.939
Confounds			
Female	47.8%	**48.5%**	0.570
Age	**31.7**	31.0	0.016
In USA	30.0%	**31.3%**	0.275
IDE	16.2%	**18.5%**	0.019
RAG	23.6%	**23.6%**	0.994
Community	**22.7%**	14.5%	0.000
Buttons	14.0%	**21.7%**	0.000

Table 3: Compares students with an Agent chatbot to those with a Tool chatbot. For each comparison, the larger value is bold, and the p-value reflects the statistical significance of the differences.

Chatbots are more popular in the IDE. In contrast, we find that placing the chatbot in the IDE dramatically increased both the number of students who engaged with it and the extent of their interactions. The percentage of students who sent at least one message was 31.2% for the IDE group, compared to just 5.6% of the Lessons group ($p = 0.000$). Similarly, students who used the chatbot in the IDE sent significantly more messages on average compared to students using a chatbot in the lessons interface (24.8 vs. 5.7; $p = 0.003$). These findings suggest that while placing the chatbot in the IDE significantly increases engagement with the tool, this increased usage does not translate into better or worse course outcomes.

4.3 Chatbot Personification: Agent or Tool?

H3 — Personification of AI will decrease lesson and assignment completion rates — is not supported. Contrary to our hypothesis, we find that personifying the chatbot as an Agent resulted in higher course completion rates. The average lesson completion rate was significantly higher for the Agent group than the Tool group (67.4% versus 65.7%, $p = 0.045$). Assignment completion and section attendance were both slightly higher for the Agent group but the differences were not statistically significant (60.4% vs. 58.8%; $p = 0.126$, 68.9% vs. 68.3%; $p = 0.500$). Other engagement metrics, such as forum activity and exam participation, also showed slight (not significant) increases for the Agent group. Altogether, these results suggest that personifying the chatbot as an Agent may positively influence course engagement, contrary to our expectations.

Note that the Agent and Tool groups differ in student demographics and chatbot features, which may impact our findings. To account for this, we conducted the regression analysis described in Section 4.4. Our analysis shows that Agent-like personification does not significantly correlate with lesson completion; however, age – which is significantly higher in the Agent group – positively correlates with this metric. While this may suggest a potential confounding effect in the group differences, it supports the finding that personification does not significantly impact student outcomes, contrary to our hypothesis.

Agent chatbots are more popular. We also find that personifying the chatbot as an Agent increased both the number of students who engaged with the chatbot, and the extent of their interactions. The percentage of students who sent at least one message was significantly higher in the Agent group than in the Tool group (12.2% vs. 7.7%; $p = 0.000$), and on average, these students sent more messages (17.6 vs. 12.6; $p = 0.146$). These findings suggest the Agent-like chatbot's welcoming message encourages more interaction.

4.4 Regression Analysis

To investigate the individual impact of each demographic characteristic and chatbot dimension on student performance we conducted a series of regression analyses. We specifically created a regression model for each evaluation metric (in Section 3.3) using student demographic and chatbot dimensions as features. This approach allows us to analyze the specific relationship between each feature and outcome metric, and verify the results that we found at the group level.

For each of the evaluation metrics, we chose the type of regression model based on the nature of the metric. For binary metrics (*Sent Message*, *Made Post*, and *Took Exam*), we used logistic regression. For count-based metrics (*Message Count* and *Post Count*), we used Poisson regression. For continuous metrics (*Exam Score*, *Assignment Completion*, *Lesson Completion*, and *Section Attendance*), we used Ordinary Least Squares (OLS).

We included all of the students to build each model, except when their outcome would be trivial. For chatbot-specific metrics (*Sent Message* and *Message Count*), we excluded students in the Control group. For count-based metrics (*Message Count* and *Post Count*), we excluded students who never engaged in the activity, as this is already captured in the engagement metric (e.g., *Sent Message*).

Each row of Table 4 summarizes the regression model for each evaluation metric; the values show the correlation direction ("+" or "−"), rounded coefficient, and statistical significance of each feature for that metric. We denote significance using * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$). "NA" indicates that the feature was excluded due to filtering out the Control group. Here, we summarize the significant findings for each demographic and chatbot feature:

- **Female:** Identifying as female shows a significant positive correlation with *Message Count*, *Assignment Completion*, *Lesson Completion*, *Section Attendance*, and *Post Count*.

Evaluation Metric	Female	Age	In USA	Agent	IDE	Community	RAG	Buttons	Control
Sent Message	+0.02	-0.06	-0.38 ***	+0.74 ***	+2.11 ***	-1.0 ***	-0.28	+0.79 ***	NA
Message Count	+0.2 ***	+0.07 ***	+0.17 ***	+0.15 ***	+1.48 ***	-0.35 **	+0.53 ***	-0.98 ***	NA
Assignment Completion	+0.02 *	+0.02 ***	+0.02 *	+0.01	-0.01	-0.01	-0.01	-0.03	+0.01
Lesson Completion	+0.03 **	+0.04 ***	+0.01	+0.01	-0.01	-0.01	-0.0	-0.02	+0.0
Section Attendance	+0.02 *	+0.05 ***	+0.03 **	+0.0	+0.0	-0.01	-0.0	-0.01	+0.0
Made Post	-0.02	+0.26 ***	-0.13 *	+0.0	-0.06	+0.07	-0.03	-0.03	-0.01
Post Count	+0.05 *	+0.23 ***	-0.15 ***	+0.02	+0.03	+0.13 ***	+0.07 *	-0.06	-0.08
Took Exam	+0.09	+0.17 ***	-0.23 ***	+0.02	-0.07	-0.05	-0.07	-0.08	+0.06
Exam Score	+0.03	+0.06 ***	-0.08 ***	+0.0	-0.02	-0.01	-0.01	-0.02	+0.02

Table 4: Regression model results for all evaluation metrics. The values show the correlation direction ("+" or "−"), rounded coefficient, and statistical significance of each feature for that metric. We denote significance using * ($p < 0.05$), ** ($p < 0.01$), and * ($p < 0.001$). "NA" indicates that the feature was excluded due to filtering out the Control group.**

- **Age:** Age shows a significant positive correlation with all metrics except *Sent Message*, which shows a negative (but not significant) correlation.
- **In USA:** USA residency shows a significant positive correlation with *Message Count*, *Assignment Completion*, and *Section Attendance*, and significantly negatively correlates with *Sent Message*, *Made Post*, *Post Count*, *Took Exam*, and *Exam Score*.
- **Agent:** Receiving an agent-personified chatbot shows a significant positive correlation with *Sent Message* and *Message Count*.
- **IDE:** Receiving the chatbot in the IDE shows a significant positive correlation with *Sent Message* and *Message Count*.
- **Community:** Receiving a chatbot with community functionality shows a significant negative correlation with *Sent Message* and *Message Count*, but significantly positively correlates with *Post Count*.
- **RAG:** Receiving a chatbot with RAG shows a significant positive correlation with *Message Count* and *Post Count*.
- **Buttons:** Receiving a chatbot with a buttons interface shows a significant positive correlation with *Sent Message* and a significant negative correlation with *Message Count*.
- **Control:** Being in the control group is not significantly correlated with any of the measured metrics.

These results suggest that specific chatbot design features (e.g., Agent, IDE, Community, RAG, Buttons) can substantially influence how often and in what ways students engage with the platform - but never significantly correlate with course completion metrics. The demographic factors (Female, Age, In USA) show meaningful correlations with both engagement and performance.

5 DISCUSSION

5.1 How do students use chatbots?

In this section, we further explore how students engaged with the chatbots and how these interactions influenced their learning outcomes. To reduce confounding factors, we only compare the experimental groups who received a chatbot with basic functionality. We used GPT-4o [35] to classify each student message into one of the following categories:

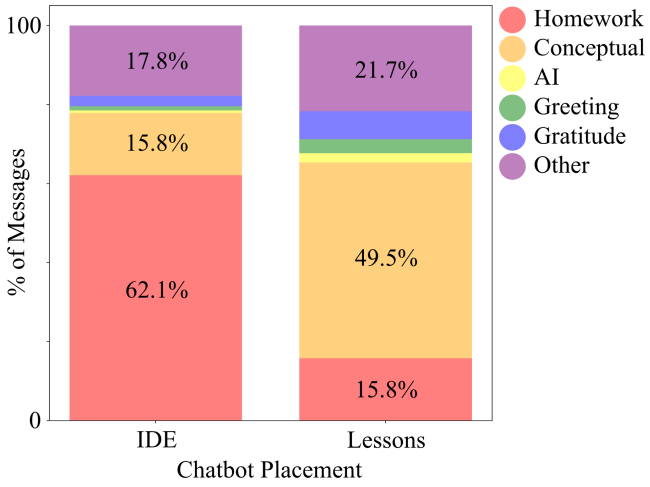

Figure 2: Distributions of the types of messages that students sent to the chatbot in the IDE and in the lessons interface. This figure includes only students who received a chatbot with Basic functionality.

(1) **Homework:** Questions about specific assignment problems (often asking for the solution).
(2) **Conceptual:** Questions about computer science concepts.
(3) **AI:** Questions about artificial intelligence (e.g., asking if the chatbot is AI).
(4) **Greeting:** Messages such as "hello" or "hi."
(5) **Gratitude:** Expressions of gratitude.
(6) **Other:** Messages that do not fit into the above categories.

Figure 2 shows dramatic difference in the distribution of student message types in the IDE versus in the lessons interface. Students using the chatbot within the IDE asked homework-related questions nearly four times more often than those who accessed the chatbot through the lessons interface (62.1% vs. 15.8%; $p = 0.000$). Conversely, students in the lessons interface were far more likely to seek conceptual help than those in the IDE (49.5% vs. 15.8%;

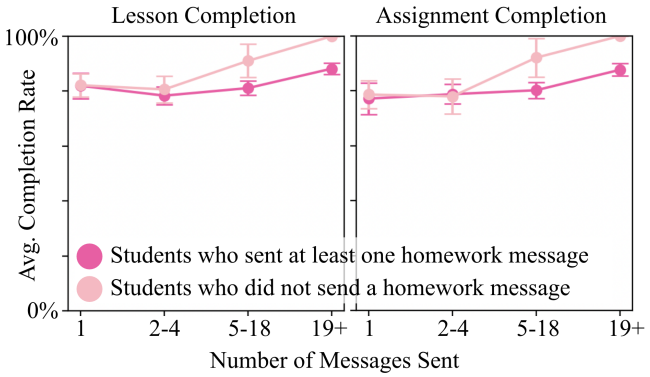

Figure 3: Compares students who sent at least one homework message, to students who did not, in terms of lesson completion and assignment completion. The error bars show the standard error of the mean.

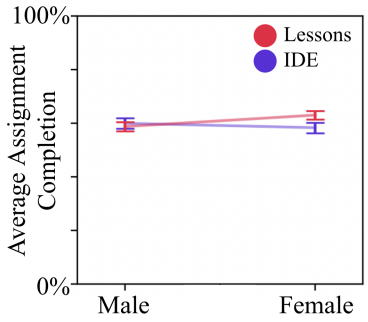

Figure 4: Assignment completion by gender, based on chatbot placement in the lessons versus the IDE. Female students complete more assignments when the chatbot is in the lessons interface, while male students complete more with IDE placement. The difference in these gender differences is significant (p = 0.045), indicating gender bias of placement.

$p = 0.000$). These findings show how dramatically placement influences how students perceive and use a tool - independent of its actual utility.

Next, we examined whether asking specific types of questions correlated with specific performance outcomes. Figure 3 compares students who sent at least one homework-related message against those who did not, grouping students by quartiles based on the total number of messages that they sent. We grouped students to reduce the confounding effect that simply sending more messages might correlate with specific performance outcomes. Our findings suggest a meaningful relationship between asking homework-related questions and lower achievement: among students who sent at least five messages, those who sent at least one homework-related message had lower lesson completion and assignment completion rates. By contrast, when we compared students who asked conceptual questions with those who did not, we found no notable differences in performance. Minimally, these results suggest that the type of questions students send are an indicator of their likeliness to complete the course.

Altogether, we see that students are particularly likely to consult a chatbot in the IDE and ask homework related questions, but that this does not improve their course outcomes. This is surprising because the chatbots were designed to be a helpful resource that assists students with their misconceptions when they are stuck; we expected the chatbot to promote student understanding with their homework, and consequently boost assignment completion. We suspect that instead, the IDE placement diverts students from solving their assignments to seeking assistance from the chatbot. This distraction might also be frustrating because the tool is designed to *not* give direct homework answers. We hypothesize that intermittent help in the IDE might be more effective and leave it to future work to explore this further.

5.2 Does chatbot impact vary by demographics?

Examining the impact of an educational intervention across student demographics is imperative for promoting equity in educational outcomes. With 6,515 students from 149 countries, we were able to examine how different chatbot experiences influenced learner performance across diverse backgrounds. Our analysis focused on three experimental conditions: *Lessons*: the students receive a chatbot integrated into lessons, *IDE*: the students receive a chatbot embedded in the IDE, and *Control*: the students do not receive a chatbot. To minimize confounding variables, we restricted our analysis to chatbots with basic functionality and measured performance using course completion metrics: assignment completion, lesson completion, and section attendance.

Gender. Due to data limitations, we restricted our gender analysis to male and female students. For each course completion metric, we calculated the difference in means across gender (male - female) and used bootstrapping to determine whether this difference varied significantly across experimental conditions. Our results show a significant difference in how chatbot placement affects assignment completion based on gender ($p = 0.045$). Specifically, women completed more assignments than men when the chatbot was in the lessons, while men completed more when the chatbot was in the IDE. Figure 4 shows this result. For all other performance metrics and comparisons, the differences across experimental conditions were not statistically significant, indicating that they did not introduce gender bias.

Human Development Index. The Human Development Index (HDI) of a country is a measure of the country's development, intended to represent the standard of living in the country [40]. To test whether the relationship between HDI and performance varied across experimental conditions, we used ordinary least squares (OLS) regression. Higher HDI consistently correlated with better performance across all three completion metrics ($p < 0.001$); however, the interaction between HDI and experimental condition was never significant ($p > 0.18$). This indicates that the relationship between HDI and performance remained stable for different chatbot placements and chatbot presence.

Age. We conducted the same analysis using student age as the predictor variable and found that, like HDI, age was a significant positive predictor of course completion across all metrics ($p < 0.028$). However, as with HDI, the interaction between age and

experimental condition was not significant, suggesting that the relationship between age and course completion did not meaningfully differ across conditions.

Our findings suggest that HDI and age are strong predictors of course completion; however, the experimental conditions did not significantly alter these relationships, meaning chatbot placement did **not** disproportionately benefit or disadvantage students based on HDI or age. This is encouraging, as these tools should be equitable and beneficial for all learners.

The only significant bias we observed was in the relationship between gender and chatbot placement for assignment completion. Specifically, chatbots integrated into lessons appeared to benefit women more, whereas chatbots in the IDE favored men. This imbalance warrants further investigation to understand its underlying causes and explore mitigation strategies.

6 LIMITATIONS AND FUTURE WORK

In retrospect, our study is limited by the number of experiment groups and the cluster randomization method we used. Because we had so many chatbot variations, our groups became relatively small and imbalanced, making it difficult to draw additional meaningful conclusions. Furthermore, counter to expectations, the chatbots in the lessons interface saw extremely low engagement rates (1.6% to 12.0%), making it even more difficult to draw conclusions about the impact of the chatbot in this context (other than the conclusion that students don't use it). If we were to do this again, we would use a different randomization strategy and fewer experiment groups, to get better balance across varying experimental conditions.

In our analysis, we utilize several metrics for course completion, site engagement, forum activity, and exam performance, and do not see a benefit of chatbots. Minimally, these metrics are an indicator of dropout, so our results suggest that the chatbots do not inspire students to stay in the course. It is important to note, however, that these signals do not encompass all possible effects on students. Future work could employ qualitative methods to understand how chatbots influence student motivation and prioritize development that inspires student curiosity.

It is also possible that the chatbots did not address any needs beyond those already met by other course resources, and thus, were not helpful in this context. The studied course consists of lessons, assignments, discussion sections, discussion forums, and additional resources, that were all carefully designed to effectively teach students how to code. We also randomly surveyed 995 students and of the 116 responses, 37% reported that they used OpenAI's ChatGPT; this suggests that a non-negligible proportion of the students used ChatGPT for help with the course. We speculate that AI-based chatbots could be more helpful in other contexts, where students have more targeted learning goals and fewer structured resources.

7 CONCLUSION

We found unexpected results in a large-scale RCT that examines how chatbot integration impacts students. We hypothesized that having a chatbot would increase course completion, that placing chatbots in the IDE would boost assignment completion, and that personifying chatbots would discourage students from using them. None of these predictions held true. Despite substantial differences in engagement based on chatbot design, we found no significant improvement in course outcomes, regardless of how the chatbot was implemented. The most significant outcomes from our study were that chatbot placement and personification dramatically affect usage, to the extent that these factors overshadow variations in functionality. These unexpected results underscore the need for continued experimentation; we believe future research in this direction will promote equitable access to meaningful learning experiences.

8 ACKNOWLEDGEMENTS

We would like to thank the Carina Foundation and all of the wonderful people of Code in Place for making this work possible. We also thank Mehran Sahami, Julie Zelenski, Juliette Woodrow, Brahm Capoor, and Ritika Kacholia for their invaluable guidance and contributions during the planning and development of this project.

REFERENCES

[1] 2024. Chat Completions API. https://platform.openai.com/docs/guides/text-generation/chat-completions-api. Accessed: 2024-06-18.
[2] 2024. Code in Place. https://codeinplace.stanford.edu/ [Online; accessed August-2024].
[3] 2024. Coursebox AI Chatbot Tutor. https://www.coursebox.ai/ai-chatbot-tutor. Accessed: 2024-07-19.
[4] 2024. Khanmigo: Khan Academy's AI-powered teaching assistant tutor. https://www.khanmigo.ai/. Accessed: 2024-07-08.
[5] 2024. Kira Learning: The AI Platform for Schools. https://www.kira-learning.com/. Accessed: 2024-07-08.
[6] 2024. Socratic: Get unstuck. Learn better. https://socratic.org/. Accessed: 2024-07-19.
[7] Mohammad Abolnejadian, Sharareh Alipour, and Kamyar Taeb. 2024. Leveraging ChatGPT for Adaptive Learning through Personalized Prompt-based Instruction: A CS1 Education Case Study. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.
[8] Patrick Bassner, Eduard Frankford, and Stephan Krusche. 2024. Iris: An AI-Driven Virtual Tutor for Computer Science Education. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1 (ITiCSE 2024)*. ACM. https://doi.org/10.1145/3649217.3653543
[9] Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Ozge Kabakci, and Rei Mariman. 2024. Generative AI Can Harm Learning. (2024). https://doi.org/10.2139/ssrn.4895486
[10] Jose Berengueres. 2024. Teaching OS315 with Chatbots: Instructional Design, Deployment, Cost & Student Attitudes. *Deployment, Cost & Student Attitudes (March 19, 2024)* (2024).
[11] Emmanuel G. Blanchard and Phaedra Mohammed. 2024. On Cultural Intelligence in LLM-Based Chatbots: Implications for Artificial Intelligence in Education. In *Artificial Intelligence in Education*, Andrew M. Olney, Irene-Angelica Chounta, Zitao Liu, Olga C. Santos, and Ig Ibert Bittencourt (Eds.). Springer Nature Switzerland, Cham, 439–453.
[12] Adelson de Araujo, Pantelis M Papadopoulos, Susan McKenney, and Ton de Jong. 2023. Supporting Collaborative Online Science Education with a Transferable and Configurable Conversational Agent. In *15th International Conference on Computer-Supported Collaborative Learning (CSCL)*.
[13] Paul Denny, Brett A Becker, Juho Leinonen, and James Prather. 2023. Chat Overflow: Artificially Intelligent Models for Computing Education-renAIssance or apocAIypse?. In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*. 3–4.
[14] Paul Denny, Stephen MacNeil, Jaromir Savelka, Leo Porter, and Andrew Luxton-Reilly. 2024. Desirable Characteristics for AI Teaching Assistants in Programming Education. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*. 408–414.
[15] Paul Denny, James Prather, Brett A Becker, James Finnie-Ansley, Arto Hellas, Juho Leinonen, Andrew Luxton-Reilly, Brent N Reeves, Eddie Antonio Santos, and Sami Sarsa. 2024. Computing education in the era of generative AI. *Commun. ACM* 67, 2 (2024), 56–67.
[16] James Finnie-Ansley, Paul Denny, Brett A Becker, Andrew Luxton-Reilly, and James Prather. 2022. The robots are coming: Exploring the implications of openai codex on introductory programming. In *Proceedings of the 24th Australasian Computing Education Conference*. 10–19.
[17] Aashish Ghimire and John Edwards. 2024. Coding with AI: How Are Tools Like ChatGPT Being Used by Students in Foundational Programming Courses. In *Artificial Intelligence in Education*, Andrew M. Olney, Irene-Angelica Chounta, Zitao

Liu, Olga C. Santos, and Ig Ibert Bittencourt (Eds.). Springer Nature Switzerland, Cham, 259–267.

[18] Simone Grassini. 2023. Shaping the future of education: exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences* 13, 7 (2023), 692.

[19] Aaron Haim, Eamon Worden, and Neil T. Heffernan. 2024. The Effectiveness of AI Generated, On-Demand Assistance Within Online Learning Platforms. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, Andrew M. Olney, Irene-Angelica Chounta, Zitao Liu, Olga C. Santos, and Ig Ibert Bittencourt (Eds.). Springer Nature Switzerland, Cham, 369–374.

[20] Michael Hammond. 2017. Online collaboration and cooperation: The recurring importance of evidence, rationale and viability. *Education and Information Technologies* 22 (2017), 1005–1024.

[21] Arto Hellas, Juho Leinonen, Sami Sarsa, Charles Koutcheme, Lilja Kujanpää, and Juha Sorva. 2023. Exploring the responses of large language models to beginner programmers' help requests. In *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 1*. 93–105.

[22] David Joyner, Zoey Anne Beda, Michael Cohen, Melanie Duffin, Amy Garcia Fernandez, Liz Hayes-Golding, Jonathan Hildreth, Alex Houk, Rebecca Johnson, Kayla Matchek, and Ana Santos. 2024. When Chatting Isn't Cheating: Mining and Evaluating Student Use of Chatbots and Other Resources During Open-Internet Exams. In *Proceedings of the 17th International Conference on Educational Data Mining*, Benjamin PaaÃŸen and Carrie Demmans Epp (Eds.). International Educational Data Mining Society, Atlanta, Georgia, USA, 143–156. https://doi.org/10.5281/zenodo.12729788

[23] Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.

[24] Muhammad Fawad Akbar Khan, Max Ramsdell, Erik Falor, and Hamid Karimi. 2023. Assessing the Promise and Pitfalls of ChatGPT for Automated Code Generation. arXiv:2311.02640 [cs.SE] https://arxiv.org/abs/2311.02640

[25] Bijan Khosrawi-Rad, Heidi Rinn, Ricarda Schlimbach, Pia Gebbing, Xingyue Yang, Christoph Lattemann, Daniel Markgraf, and Susanne Robra-Bissantz. 2022. Conversational agents in education–a systematic literature review. (2022).

[26] Mark Liffiton, Brad E Sheese, Jaromir Savelka, and Paul Denny. 2023. Codehelp: Using large language models with guardrails for scalable support in programming classes. In *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research*. 1–11.

[27] Rongxin Liu, Carter Zenke, Charlie Liu, Andrew Holmes, Patrick Thornton, and David J. Malan. 2024. 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* (Portland, OR, USA) *(SIGCSE 2024)*. Association for Computing Machinery, New York, NY, USA, 750–756. https://doi.org/10.1145/3626252.3630938

[28] Rose Luckin and Wayne Holmes. 2016. Intelligence unleashed: An argument for AI in education. (2016).

[29] Wenhan Lyu, Yimeng Wang, Tingting (Rachel) Chung, Yifan Sun, and Yixuan Zhang. 2024. 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 (L@S '24)*. ACM, 63–74. https://doi.org/10.1145/3657604.3662036

[30] Boxaun Ma, Li Chen, and Shin'ichi Konomi. 2024. Enhancing Programming Education with ChatGPT: A Case Study on Student Perceptions and Interactions in a Python Course. arXiv:2403.15472 [cs.CY] https://arxiv.org/abs/2403.15472

[31] Ali Malik, Juliette Woodrow, Brahm Capoor, Thomas Jefferson, Miranda Li, Sierra Wang, Patricia Wei, Dora Demszky, Jennifer Langer-Osuna, Julie Zelenski, Mehran Sahami, and Chris Piech. 2023. Code in Place 2023: Understanding learning and teaching at scale through a massive global classroom. https://piechlab.stanford.edu/assets/papers/codeinplace2023.pdf.

[32] Silvia Milano, Joshua A McGrane, and Sabina Leonelli. 2023. Large language models challenge the future of higher education. *Nature Machine Intelligence* 5, 4 (2023), 333–334.

[33] Chancharik Mitra, Mihran Miroyan, Rishi Jain, Vedant Kumud, Gireeja Ranade, and Narges Norouzi. 2024. Elevating Learning Experiences: Leveraging Large Language Models as Student-Facing Assistants in Discussion Forums. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2*. 1752–1753.

[34] Allen Nie, Yash Chandak, Miroslav Suzara, Malik Ali, Juliette Woodrow, Matt Peng, Mehran Sahami, Emma Brunskill, and Chris Piech. 2024. The GPT Surprise: Offering Large Language Model Chat in a Massive Coding Class Reduced Engagement but Increased Adopters Exam Performances. https://doi.org/10.31219/osf.io/qy8zd

[35] OpenAI. 2024. OpenAI Model Documentation. https://platform.openai.com/docs/models Accessed: 2024-02-04.

[36] Zachary A Pardos and Shreya Bhandari. 2023. Learning gain differences between ChatGPT and human tutor generated algebra hints. *arXiv preprint arXiv:2302.06871* (2023).

[37] Chris Piech. [n. d.]. Bootstrapping. https://chrispiech.github.io/probabilityForComputerScientists/en/part4/bootstrapping/. Accessed: 2025-01-03.

[38] Christopher Piech, Ali Malik, Kylie Jue, and Mehran Sahami. 2021. Code in place: Online section leading for scalable human-centered learning. In *Proceedings of the 52nd acm technical symposium on computer science education*. 973–979.

[39] James Prather, Brent Reeves, Juho Leinonen, Stephen MacNeil, Arisoa S. Randrianasolo, Brett Becker, Bailey Kimmel, Jared Wright, and Ben Briggs. 2024. The Widening Gap: The Benefits and Harms of Generative AI for Novice Programmers. arXiv:2405.17739 [cs.AI] https://arxiv.org/abs/2405.17739

[40] United Nations Development Programme. n.d.. Human Development Index (HDI). https://hdr.undp.org/data-center/human-development-index#/indicies/HDI Accessed: [date].

[41] Laryn Qi, JD Zamfirescu-Pereira, Taehan Kim, Björn Hartmann, John DeNero, and Narges Norouzi. 2024. A Knowledge-Component-Based Methodology for Evaluating AI Assistants. *arXiv preprint arXiv:2406.05603* (2024).

[42] Ruan Reis, Gustavo Soares, Melina Mongiovi, and Wilkerson L Andrade. 2019. Evaluating feedback tools in introductory programming classes. In *2019 IEEE Frontiers in Education Conference (FIE)*. IEEE, 1–7.

[43] Sherry Ruan, Allen Nie, William Steenbergen, Jiayu He, JQ Zhang, Meng Guo, Yao Liu, Kyle Dang Nguyen, Catherine Y Wang, Rui Ying, James A Landay, and Emma Brunskill. 2023. Reinforcement Learning Tutor Better Supported Lower Performers in a Math Task. arXiv:2304.04933 [cs.AI] https://arxiv.org/abs/2304.04933

[44] 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).

[45] Brad Sheese, Mark Liffiton, Jaromir Savelka, and Paul Denny. 2024. Patterns of student help-seeking when using a large language model-powered programming assistant. In *Proceedings of the 26th Australasian Computing Education Conference*. 49–57.

[46] Ben Arie Tanay, Lexy Arinze, Siddhant S Joshi, Kirsten A Davis, and James C Davis. 2024. An Exploratory Study on Upper-Level Computing Students' Use of Large Language Models as Tools in a Semester-Long Project. *arXiv preprint arXiv:2403.18679* (2024).

[47] Karan Taneja, Pratyusha Maiti, Sandeep Kakar, Pranav Guruprasad, Sanjeev Rao, and Ashok K. Goel. 2024. Jill Watson: A Virtual Teaching Assistant powered by ChatGPT. arXiv:2405.11070 [cs.AI] https://arxiv.org/abs/2405.11070

[48] S. Tegos, S. Demetriadis, and P.M. et al. Papadopoulos. 2016. Conversational agents for academically productive talk: a comparison of directed and undirected agent interventions. *Intern. J. Comput.-Support. Collab. Learn* 11 (2016), 417–440.

[49] Stéphan Vincent-Lancrin and Reyer Van der Vlies. 2020. Trustworthy artificial intelligence (AI) in education: Promises and challenges. (2020).

[50] Sierra Wang, John Mitchell, and Chris Piech. 2024. A large scale RCT on effective error messages in CS1. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*. 1395–1401.

[51] Juliette Woodrow, Ali Malik, and Chris Piech. 2024. Ai teaches the art of elegant coding: Timely, fair, and helpful style feedback in a global course. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*. 1442–1448.

[52] Mike Wu, Noah Goodman, Chris Piech, and Chelsea Finn. 2021. Prototransformer: A meta-learning approach to providing student feedback. *arXiv preprint arXiv:2107.14035* (2021).

[53] J. D. Zamfirescu-Pereira, Laryn Qi, Björn Hartmann, John DeNero, and Narges Norouzi. 2024. 61A-Bot: AI homework assistance in CS1 is fast and cheap – but is it helpful? arXiv:2406.05600 [cs.CY] https://arxiv.org/abs/2406.05600