Understanding Programming Students' Help-Seeking in the Era of Generative AI

Jacob Penney
jmp458@nau
Northern Arizona University
Flagstaff, AZ, USA

Pawan Acharya
pa577@nau.edu
Northern Arizona University
Flagstaff, AZ, USA

Peter Hilbert
peh53@nau.edu
Northern Arizona University
Flagstaff, AZ, USA

Priyanka Parekh
Priyanka.Parekh@nau.edu
Northern Arizona University
Flagstaff, AZ, USA

Anita Sarma
anita.sarma@oregonstate.edu
Oregon State University
Corvallis, OR, USA

Igor Steinmacher
igor.steinmacher@nau.edu
Northern Arizona University
Flagstaff, AZ, USA

Marco A. Gerosa
marco.gerosa@nau.edu
Northern Arizona University
Flagstaff, AZ, USA

Abstract

Novice programming students frequently engage in help-seeking to find information and learn about programming concepts. Among the available resources, generative AI (GenAI) chatbots appear resourceful, widely accessible, and less intimidating than human tutors. Programming instructors are actively integrating these tools into classrooms. However, our understanding of how novice programming students trust GenAI chatbots—and the factors influencing their usage—remains limited. To address this gap, we investigated the learning resource selection process of 20 novice programming students tasked with studying a programming topic. We split our participants into two groups: one using ChatGPT (n=10) and the other using a human tutor via Discord (n=10). We found that participants held strong positive perceptions of ChatGPT's speed and convenience but were wary of its inconsistent accuracy, making them reluctant to rely on it for learning entirely new topics. Accordingly, they generally preferred more trustworthy resources for learning (e.g., instructors, tutors), preferring ChatGPT for low-stakes situations or more introductory and common topics. We conclude by offering guidance to instructors on integrating LLM-based chatbots into their curricula—emphasizing verification and situational use—and to developers on designing chatbots that better address novices' trust and reliability concerns.

CCS Concepts

• **Social and professional topics** → **CS1**; • **Human-centered computing** → **Empirical studies in HCI**; • **Applied computing** → *Interactive learning environments.*

CompEd 2025, Gaborone, Botswana

ACM ISBN 979-8-4007-1929-5/2025/10
https://doi.org/10.1145/3736181.3747165

Keywords

CS1, novice programming students, help-seeking behavior, generative AI, human–AI interaction, computing education

ACM Reference Format:
Jacob Penney, Pawan Acharya, Peter Hilbert, Priyanka Parekh, Anita Sarma, Igor Steinmacher, and Marco A. Gerosa. 2025. Understanding Programming Students' Help-Seeking in the Era of Generative AI . In *Proceedings of the ACM Global Computing Education Conference 2025 Vol 1 (CompEd 2025), October 21–25, 2025, Gaborone, Botswana.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3736181.3747165

1 Introduction

Students learning to program often address knowledge gaps through help-seeking [19], weighing factors [35] to select from resources such as web search [41], programming Q&A sites [6], video platforms [4, 27], and traditional options like teaching assistants [28], books, peers, and instructors [5].

Generative AI conversational interfaces (GenAI chatbots) have emerged as powerful, accessible resources that generate both excitement and concern in programming education. Instructors worry that such tools may miseducate students if over-trusted [37, 42], since they can produce inaccurate or false information [15] and lack transparency [14]. At the same time, GenAI chatbots have shown the ability to solve complex programming problems [7], provide on-demand guidance [24], and offer free or low-cost access. A growing body of literature explores how programming students use them as learning resources [13].

Despite these advancements, little is known about how novice programming students trust and use GenAI chatbots as learning resources. This study investigates the factors shaping novices' resource usage in the context of GenAI chatbots, posing the research question: **What factors shape how novice CS students use Generative AI to learn programming concepts?**

We conducted an exploratory study with 20 novice students, assigning them to learn a programming topic with ChatGPT (n=10) or a human tutor via text chat (n=10). By isolating each resource, we could qualitatively compare engagement, trust, and cost–benefit

reasoning [29]. Participants completed think-aloud tasks and debriefing interviews, allowing us to examine their decisions about when and how to use ChatGPT or a human tutor.

Our findings show that students' perceptions of ChatGPT's strengths and weaknesses influenced its place in their learning process. While participants valued ChatGPT's speed and accessibility, they were reluctant to rely on it for learning new topics, generally preferring more trustworthy resources. We conclude with guidance for instructors integrating chatbots into CS education and recommendations for developers to improve trust and reliability.

2 Related Work

Research shows that large language models (LLMs), such as those used by ChatGPT, demonstrate near-human performance for some tutoring tasks in computer science but struggle in others. Phung et al. [33] found that LLMs approach human tutors in program repair and contextualized explanation, yet lag in grading feedback or synthesizing new tasks. Jury et al. [16] showed that LLM-generated worked examples feature meaningful step-by-step logic valued by both experts and CS1 students. Leinonen et al. [23] reported that novices perceive ChatGPT's code explanations as clearer than peers', prompting more help-seeking (e.g., examples and reasoning). Balse et al. [1] observed that LLM explanations of code with logical errors were indistinguishable from student-created ones, yet identified at least one mistake in 93% of cases. Hoq et al. [12] showed that machine learning classifiers can detect ChatGPT-generated code submissions with over 90% accuracy.

As these tools enter programming education, research increasingly considers their benefits, drawbacks, and impact on instructors and students. Prather et al. [34] noted that while instructors worry about academic dishonesty and "genuine learning", they also see higher-quality student work and reduced office-hour demand. This tension underscores the need to study GenAI's role in computing education.

2.1 The Cost-Value Framework of Help-Seeking

The Cost-Value Framework explains how students weigh social and cognitive factors when seeking help. Nelson-Le Gall [29] proposed that learners balance costs (e.g., embarrassment, effort) and benefits (e.g., understanding, task success). Karabenick and Berger [18] extended this model by emphasizing learners' goals, self-regulation, and context. Newman distinguishes *instrumental* help-seeking for deeper learning from *executive* help-seeking for quick solutions [30]. Ryan and Pintrich [36] showed that performance-oriented learners often bypass higher-cost resources (e.g., instructors) for low-stakes questions. Karabenick [17] identified trust, instructor approachability, and task complexity as core drivers, and Makara and Karabenick [26] showed how computer-mediated communication can lower social risk and accelerate responses.

This framework suits our study because it accounts for the interplay between cognitive factors (e.g., topic difficulty, prior knowledge) and social dynamics (e.g., tutor–student relationships, perceptions) in tutoring interactions. Its emphasis on context aligns with our investigation of how different tutoring approaches affect help-seeking, and its focus on trust and psychological safety informs our analysis of why students might choose AI chatbots over human tutors.

2.2 Novice Programming Student Help-Seeking with Generative AI in Higher Education

Recent studies examine how students approach GenAI tools for learning. Hou et al. [13] found that while students value ChatGPT's efficiency, they still favor traditional resources, prioritizing accuracy and trusting ChatGPT for low-pressure queries. Keuning et al. [20] reported that less experienced students are less likely to use GenAI, and more advanced students worry about over-reliance. Haindl and Weinberger [9] found that part-time undergraduates view ChatGPT as suitable for learning programming concepts, while Skripchuk et al. [38] noted that novices often prefer web search over GenAI unless they feel in control of AI output.

Our prior work examined conversational agent design and interaction patterns in programming education. [32] identified instructors' expectations for LLM-based tutors, including stepwise guidance, adaptive explanations, and pedagogical focus. In a lab study, [31] found that ChatGPT users favored brief prompts yielding long replies, whereas human tutors offered more contextualized explanations. These studies show how tool affordances and social dynamics shape help-seeking.

A gap remains in understanding the contextual factors influencing when and why novices use GenAI for learning. Our study addresses this gap by directly comparing GenAI and human tutors, using think-aloud protocols and interviews to provide a fuller picture of students' resource hierarchies, preferences, and trust thresholds—insights useful to educators designing effective learning environments and developers building pedagogically sound AI tools.

3 Research Design

To investigate the factors influencing novice CS students' help-seeking decisions, we designed an exploratory qualitative laboratory study in which participants learned a programming concept using either ChatGPT or a human tutor via a popular instant messaging platform, Discord. We adopted a qualitative approach to uncover decision-making factors and contextual preferences that shape resource selection. This methodological choice aligns with similar small-sample, in-depth studies (e.g., [13]) that focus on rich participant reflections and emergent themes. Controlling the tutor variable allowed us to isolate the influence of each type of tutor on participants' help-seeking behaviors, and perceptions of trust, costs, and benefits—exploring help-seeking behaviors specific to each type of assistance. Additionally, by giving all participants to work with the same concept and task, we controlled the difficulty and knowledge requirements.

We received IRB approval for this research design through our institution's reviewing board. Following our IRB protocol, all published results are presented in aggregated form with personally identifiable information removed. To support reproducibility and transparency in our research, all study artifacts are included in the artifact package [1].

[1] https://osf.io/7uxdz/?view_only=35e7f8a8fcb345839bcd8409c2ec660e

Table 1: Participant demographics

ID	Level	Major	Course	Age	Gender	Group
P1	Ugrad	CS	Intro	18–20	M	GPT
P2	Ugrad	English	Intro	21–30	M	GPT
P3	Ugrad	Informatics	Intro	18–20	W	GPT
P4	Ugrad	Comp Eng	Intro	21–30	M	GPT
P5	Grad	CS	DSA	21–30	M	GPT
P6	Grad	CS	DSA	21–30	M	GPT
P7	Grad	CS	DSA	21–30	M	GPT
P8	Grad	CS	DSA	21–30	M	GPT
P9	Grad	CS	DSA	21–30	M	GPT
P10	Grad	CS	DSA	21–30	M	GPT
P11	Grad	CS	DSA	21–30	M	Discord
P12	Grad	CS	DSA	21–30	M	Discord
P13	Grad	CS	DSA	21–30	M	Discord
P14	Grad	CS	DSA	31–40	M	Discord
P15	Ugrad	CS	Intro	18–20	M	Discord
P16	Grad	CS	DSA	21–30	M	Discord
P17	Grad	CS	DSA	21–30	M	Discord
P18	Grad	CS	DSA	21–30	W	Discord
P19	Grad	CS	DSA	21–30	M	Discord
P20	Grad	CS	DSA	21–30	W	Discord

3.1 Recruitment

We used volunteer and snowball sampling to recruit 20 participants from a pool of 340 students in nine sections of three introductory programming courses at an American public university between February and November 2024. We advertised through two channels: (1) distributing study flyers via email through course instructors and (2) making short presentations in class. A total of 40 students completed an initial screening survey regarding prior experiences with and attitudes toward chatbots in education; five were dismissed because they did not match the course/expertise requirements. We then invited the remaining 35 eligible students for 30–60 minute lab sessions over Zoom. 20 of the remaining participants responded, with whom we scheduled sessions. Participant demographics are presented in Table **??**. We include these graduate students as "introductory" because they were enrolled in entry-level programming courses designed for students without formal programming education or who require remedial programming education.

3.2 Lab Study

For each session of the study, a proctor introduced each task, then disabled their camera and microphone but remained available for clarifying questions. Participants had five minutes for Phases 1 and 3 and ten minutes for Phase 2. Clarifying questions or requests for technical assistance did not reduce that time. They could not request assistance with the quizzes or revisit prior phases, and they were not told details about the contents of the phases ahead of time. We asked participants to engage in a think-aloud protocol during each phase to capture their experience as it happened.

Phase 1: Pre-assessment – Participants completed a custom quiz about pointers in the C language and a brief sentiment questionnaire, allowing us to capture their initial competency with this foundational topic and their sentiments about and trust in LLM-based chatbots. The quiz was custom-made to align with CS1 learning outcomes at our institution and consisted of three multiple-choice (MC) questions and one long answer (LA) question. The quiz questions are: 1) *What is a pointer?* (MC), 2) *Which of the following best describes why pointers are used in programming?* (MC), 3) *What does dereferencing a pointer mean?* (MC), and 4) *Describe risk(s) associated with using pointers, such as what the problem is and what dangers it poses.* (LA). The full quiz and sentiment survey can be found in the paper's artifact package.[2].

Phase 2: Learning through the chat interface – All participants studied the same concept, either with ChatGPT (GPT-3.5) or with a human tutor via Discord. The human tutor was described only as someone with formal CS education. We refer to these groups as "ChatGPT" and "Discord", respectively. Participants were not limited in the questions they could ask their assigned tutor, but none were not permitted to revisit Phase 1 to view the quiz questions.

Phase 3: Post-assessment – We repeated the same quiz and sentiment questionnaire to capture changes in participants' knowledge and attitudes. This duplication was not disclosed beforehand to mitigate bias.

Phase 4: Debrief – We asked three questions to all participants: (1) "Can you describe how it felt using the digital tutor to learn the given computer science concept?", (2) "To what extent did you trust the tutor?", and (3) "Could you see yourself using this method for learning new concepts in the future?". We also adapted follow-up questions to the standard questions based on individual actions or responses. Participant responses to these questions were open-ended, with any length being acceptable and no time limit. All interviews lasted between 10 and 25 minutes.

3.3 Data Analysis

To analyze the think-aloud and debriefing data, the primary researcher iteratively conducted open and axial coding. First, the research team transcribed recorded content. Then, the primary researcher conducted the first round of coding, during which they read the text, identified quotes with relationships to our research question, and labeled the quotes with primary and secondary codes, intended to capture their context and intent. For example, some codes relevant to the results presented here are *would not use ChatGPT to learn new topics*, *might use ChatGPT among many tools*, *prefers digital learning resources to avoid embarrassment*, *cannot verify ChatGPT's accuracy*.

After completing the initial coding, the research team convened to deliberate over, rewrite, and reorganize the researcher's codes. The researcher integrated these edits and continued with rounds of coding. In all, the research team met four times, each for two to three hours, to deliberate codes. After reaching stability and code saturation [11], the research team stopped meeting, and the primary researcher analyzed the rest of the debriefing data. The resulting codebook is presented in the artifact package.[3].

4 Results

In this section, we present our results organized by the major themes that emerged from the qualitative analysis of the think-aloud and debrief sessions.

ChatGPT is one option in a hierarchy rather than a universal solution for learning. Help-seeking behavior in both settings

[2]https://osf.io/7uxdz/?view_only=35e7f8a8fcb345839bcd8409c2ec660e
[3]https://osf.io/7uxdz/?view_only=35e7f8a8fcb345839bcd8409c2ec660e

Table 2: Resources participants use for learning

Resource	Participants	Total
Tutors	P11, P13, P15, P20	4
Instructors	P3, P4, P5, P9	4
Web search	P3, P4, P6, P7	4
Video tutorials	P5, P15, P19	3
Online courses	P5, P19	2
Mentors	P3	1
Online documentation	P7	1
Class modules	P10	1

Table 3: Factors influencing resource selection

Factor	Participants	Total
Reliability of resource	P1, P3, P5, P9, P10, P14	6
Trust required for question	P1, P2, P3, P4, P6	5
Ease of use	P1, P2, P6, P17, P20	5
Clarity of explanations	P12, P14, P15, P16	4
Topic/domain of question	P1, P2, P3, P6, P7	5
Convenience	P3, P5, P7, P20	4
Prior experience with topic	P2, P5, P7, P10	4
Social cost (e.g., embarrassment)	P3, P5, P11	3
Ease of accessing resource	P5, P6	2
Method used to teach	P2, P13	2
Specific to niche context	P6	1
Fun factor	P4	1

involved constructing a hierarchy of available resources rather than relying on a single universal solution. Our participants perceived that their help-seeking strategy they use depended on factors such as the nature of the task and the suitability of the resource for that task (Table 2 and Table 3). P3 highlighted that trust in a resource influences its place in their help-seeking hierarch: "the topic is always like the big part of where whether...[they] trust it or not". Similarly, P4 noted that ChatGPT serves as "a really good ice breaker for topics", suggesting that it plays a role in initiating understanding rather than serving as a comprehensive source. Importantly, resources were not seen as interchangeable but as complementary within a structured approach to help-seeking. As P10 explained, "I will not use ChatGPT completely. So, I will use that up to some extent. Then I will use the modules or textbooks to cross-refer whether the information given by ChatGPT is correct", illustrating an approach where multiple sources are consulted to verify and refine understanding.

Participants have positive views of ChatGPT's performance. Concerning its positive qualities as a tutor, students found ChatGPT to be fast (n=7), accessible (n=1), easy to use (n=4), and capable of producing convenient results (n=3). P6 said *"...through Google searches or those things, we may not get appropriate answers. ChatGPT almost instantly gives the data we need for our specific problems"*. Some Discord group participants (n=4) reported that ChatGPT is faster than the human tutor they interacted with based on past experiences with ChatGPT. Some ChatGPT group participants (n=2) appreciated that ChatGPT removed the need to compete with other students for resources or to burden others with questions: *"It's also helpful that you can ask for as many different examples and build upon your question as many times as you need. And it'll be just a thousand times faster"* [P3].

Table 4: Positive sentiments regarding trust with ChatGPT

Opinion/Sentiment	Participants	Total
Reliable for familiar tasks/topics	P1, P2, P5, P7	4
Reliable as an introduction to topics or terminology	P3, P4	2
Trustworthy with STEM topics or math	P3	1
Trustworthy for programming concepts	P2	1
Reliable for introductory programming	P5	1

Table 5: Negative sentiments regarding trust with ChatGPT

Opinion/Sentiment	Participants	Total
Output is unreliable, challenging to verify	P1, P3, P5, P9, P10	5
Untrustworthy with calculations	P1, P6	2
Outputs need to be verified with other resources	P3, P10	2
Unreliable for complex topics	P5, P17	2
Issues with even basic programming questions	P6	1
Mixed opinions about accuracy for definitions	P3	1
Only useful as a last-ditch effort	P3	1

Table 6: Sentiments regarding specific ChatGPT use cases

Sentiment	Participants	Total
Wouldn't use for new topics	P2, P3, P5, P7, P9, P18	6
More useful with past experience with subject	P1, P6, P7, P10	4
Useful as an initial resource for exploratoration	P1, P4, P6	3
Prefer instructors/tutors before ChatGPT	P3, P7, P18	3
Would use ChatGPT for new topics	P1, P11, P20	3
Wouldn't use ChatGPT by itself	P10, P12	2

Participants do not trust ChatGPT for learning. Our participants expressed that ChatGPT's untrustworthiness has significant weight compared to the positive facets of its user experience. Students see ChatGPT as producing output with inconsistent veracity (n=5), which they have little ability to detect and have to verify using external resources (n=2). For example, P3 said *"...the only thing that... I feel like would help build trust is just making sure that that information is as genuine as it could be"*.

Participants would not depend on ChatGPT to learn new topics. Participants' distrust in GenAI is highly contextualized by what they are trying to learn. As shown in Table 6, most participants in the ChatGPT group said that they would not use ChatGPT to learn a new topic (n=6) or would not use it without also using other resources (n=2). Despite the Discord group being provided only with the information that the tutor was "someone with an academic background in computer science and a history of tutoring", half of Discord participants (P11 through P15) indicated that they trusted the digital tutor enough to learn new topics with them and one indicated 100% trust in the tutor: *"if I don't know anything about the concept and I'm asking the question, yeah, whatever he's saying, he's correct. He's rightful to me."* Related, our participants perceived that they would have an easier time verifying ChatGPT's outputs if they were using it for topics they have familiarity with: *"you need knowledge so if you are... well versed with the topic before then you can exactly understand if it's giving you the right thing or not"* [P5].Still, P15 said they would not mind foregoing quick responses in favor of more accurate information.

Participants view ChatGPT as more useful in low stakes and lower effort phases of learning. Students had a relatively

Table 7: Issues With Digital Tutors

Issue cited	Participants	Total
Human tutors respond slowly	P15, P16, P17, P20	4
Lack of multimedia tools	P2, P18, P20	3
Lack of interactive methods	P2, P18	2
Typing-related delays	P19	1
Language barriers	P18	1
Use of technical jargon	P15	1

positive view of using ChatGPT during an initial information-seeking phase because it is accessible, works quickly, returns concentrated results, has a low social cost such that they can ask questions they may be embarrassed to ask a human, and the trust required is commensurate with the low stakes. This may include finding introductions to topics or definitions and terminology, a lead they can pursue deeper through a web search, or as a way to get actionable explanations for foundational concepts that they will then practice on their own, as said P11: *"...I will use [ChatGPT] up to some extent, then I will use the modules or textbooks to cross-refer whether the information given by ChatGPT is correct or not"*. On the other hand, some students view that this may misguide them right away and will consult ChatGPT as a last resort, first preferring instructors and perhaps web search after: *"I would usually do a Google search first and just compare different resources, and looking for like more credible sources... before I use ChatGPT... and I would, before a simple Google search, I'd always go to like a professor..."* [P3].

Students prefer the pedagogical experience they receive in class. Overall, our participants saw humans as more reliable, better at explanations, and more trustworthy. Students want digital tutors to deliver high-quality and trustworthy knowledge on the subject matter using similar methods and resources as their instructors and in-person tutors (see Table 7). For example, P2 felt that how fast ChatGPT returns content is not interesting because it has no bearing on the speed at which one learns: *"When I asked a question to ChatGPT... it just shows me like all of the important aspects of pointers in C language... instead of like we did in the class, we go deeper. We go from the superficial stuff and then we go deeper. So we can learn it more quick"*. On the other hand, participants value the human qualities of communication with instructors and tutors, citing emotional connection, physical cues, and experience with a subject as important benefits over GenAI chatbots: *"We ask the question and we are making eye contact, we are able to talk to the tutor, and we are able to understand what exactly he wants to say"* [P19]. They appreciate that humans intuit their doubts and understand their questions more readily (*"It's definitely easy to communicate with the tutor, as it's understanding what I'm trying to tell"* [P18]), provide tailored responses, and ask clarifying questions when uncertain about students' questions. Students also trust human tutors to admit when they do not know an answer, which can enhance trust.

Findings

Students currently choose to use ChatGPT for low-stakes tasks that they would like to conduct quickly. They have a middling trust for ChatGPT but like its UX. Among our participants, it does not disrupt their normal learning resource selection process.

5 Discussion

This study investigated novice programming students' resource selection processes when using GenAI chatbots, such as ChatGPT, compared to human tutors. The findings reveal that while students appreciate ChatGPT's accessibility, speed, and convenience, they harbor significant concerns about its trustworthiness and role in supporting deep learning. These insights offer critical implications for CS pedagogy and the design of pedagogical GenAI tools.

The Cost-Value Framework of Help-Seeking explains Students' Motivations to Use ChatGPT for Learning. Our findings align with the Cost-Value Framework of Help-Seeking, which posits that learners balance the effort and risk of seeking help against the perceived benefit. Many participants viewed ChatGPT as a low-cost option—it is socially safe, always available, and responds quickly. However, they also viewed it as low-value in high-stakes contexts due to its inconsistency and lack of trustworthiness. In contrast, human tutors and instructors were perceived as higher-cost (slower, less convenient, sometimes intimidating) but also higher-value, particularly for complex or unfamiliar topics. This cost-value reasoning was apparent in how students assembled resource hierarchies, where ChatGPT occupied a useful but limited role—often as a preliminary tool for familiar or low-stakes content. Framing their decisions through this lens helps explain why students continue to prefer traditional resources for deeper learning and reinforces the importance of scaffolding students' help-seeking strategies in GenAI-integrated classrooms.

Balancing Strengths and Limitations of ChatGPT in Education. Our results indicate that ChatGPT is most useful for low-stakes tasks, such as gathering introductory information or exploring definitions. Participants praised its ability to quickly deliver concise and targeted responses, which aligns with prior findings on the strengths of generative AI tools in supporting exploratory learning [24**?**]. This ease of use and speed make ChatGPT a valuable resource for addressing students' immediate, surface-level queries, especially when time is limited or access to human resources is unavailable.

However, our findings also highlight ChatGPT's perceived limitations in tasks requiring high trust or deeper conceptual understanding. Participants expressed concerns about the inconsistent accuracy of its outputs and their inability to verify the responses, echoing broader critiques of genAI's lack of transparency and potential for misinformation [14, 15]. These results suggest that ChatGPT should be positioned as a complementary tool in programming education, best suited for initial exploration or low-stakes information retrieval rather than as a standalone resource for learning complex concepts.

Addressing Trust and Reliability Concerns. A critical barrier to ChatGPT's broader adoption in education is the trust deficit observed among students. Our participants' skepticism mirrors existing research highlighting public distrust in generative AI systems due to their opaque reasoning processes and occasional inaccuracies [3]. While participants acknowledged ChatGPT's usefulness for

certain tasks, many relied on external resources to verify its outputs, which reduced its perceived reliability as a learning resource.

To address these concerns, developers should prioritize features that enhance trust and reliability. Providing citations for responses, flagging potentially unreliable outputs, and incorporating error-detection mechanisms could significantly improve user confidence [14]. Additionally, features that emulate human-like qualities, such as acknowledging uncertainty or asking clarifying questions, could align ChatGPT's behavior with the expectations students have of trustworthy tutors [39].

Enhancing Engagement and Pedagogical Value. Our findings also emphasize the value students place on interactive and engaging learning experiences. Participants frequently praised human tutors for their ability to adapt explanations to individual needs, scaffold complex concepts, and offer emotional support. These qualities of human instruction are well-documented in education research and contribute to students' conceptual understanding and confidence [5, 25].

Developers of GenAI tools should aim to replicate these pedagogical strategies. For instance, chatbots could incorporate step-by-step guidance, scaffolding techniques, or context-aware responses that build on users' prior knowledge. Such enhancements could make ChatGPT more effective for learning tasks beyond simple information retrieval, addressing gaps in its current utility as identified by participants.

Contextualizing ChatGPT's Role in CS Pedagogy. Our findings reinforce the importance of situational awareness when integrating ChatGPT into programming education. Students' resource selection was influenced by task complexity, prior familiarity with the subject, and the stakes involved, which aligns with research on help-seeking in academic settings [19]. While ChatGPT was valued for its speed and accessibility in low-stakes scenarios, participants preferred human instructors or tutors for high-stakes tasks or learning new concepts.

Instructors should guide students on when and how to use ChatGPT effectively, emphasizing its strengths while cautioning against over-reliance. For example, students can be encouraged to use ChatGPT as a starting point for exploring new topics but to verify its outputs through more authoritative resources, such as textbooks, peers, or instructors. This hybrid approach leverages the complementary strengths of AI tools and traditional resources, creating a more balanced and effective learning ecosystem [22**?**].

Implications for Future Research and Development. Our findings suggest that integrating ChatGPT into CS pedagogy requires careful consideration of its strengths and limitations. Future research could explore how generative AI tools impact long-term learning outcomes, particularly in developing computational thinking skills [40]. Additionally, longitudinal studies may provide insights into how trust and usage patterns evolve as students gain experience with these tools.

For developers, these results underscore the need for user-centered design approaches that address trust, transparency, and engagement. By aligning chatbot features with students' expectations and educational needs, GenAI tools can become more effective and trusted resources for novice programmers.

6 Limitations

As with any empirical study, there are limitations to consider when interpreting our results.

Our study's small sample size from a single institution limits generalizability. However, our sample aligns with typical qualitative exploratory research, which commonly includes 6 to 30 participants for thematic discovery [2, 8, 10, 11, 21]. The volunteer and snowball sampling methods may have introduced selection bias, attracting participants already interested in or experienced with AI tools. However, group assignments were randomized to prevent bias toward the treatment.

The controlled laboratory environment and limited session duration might not fully reflect natural learning conditions. Although allowing free choice of resources could reveal natural help-seeking patterns, it would introduce confounds due to varied preferences, prior experiences, and convenience. To avoid these confounds, we randomly assigned students to either ChatGPT or a human tutor, isolating each tutor's influence. Furthermore, having all participants learn the same concept (pointers in C) facilitated direct comparisons.

The use of GPT-3.5 provides a specific snapshot of generative AI capabilities; findings may differ with other AI models or future technological advancements.

Lastly, qualitative analysis inherently involves subjective interpretation. To mitigate potential biases, we employed iterative coding, utilized multiple reviewers, and conducted consensus discussions, thereby enhancing the reliability and validity of our findings.

7 Future Work

Future work will examine how pedagogically designed chatbots influence resource selection, learning outcomes, and metacognition. Building on this study, we will compare environments offering both pedagogical and non-pedagogical agents and assess outcomes across different designs.

8 Conclusion

In this research, we investigated *What factors shape how novice CS students use Generative AI to learn programming concepts?* We found that students have a robust learning resource selection process influenced by many factors and that ChatGPT does not subvert this process but is subsumed by it. We also found that novice CS students do not currently perceive ChatGPT as suitable for studying new concepts and found behaviors that an LLM-based chatbot should employ to become more useful to them for this and other learning tasks.

From these results, we offer advice to students, instructors, and chatbot developers about interaction with GenAI chatbots, students support to learn about and through such tools, and about how to develop pedagogical chatbots to appeal to instructors and support students, respectively.

Acknowledgments

This work was supported by the National Science Foundation (NSF) under Grant Nos. 2236198, 2303042, 2235601, and 2303043. We thank the students who participated in this study for allowing us to observe them.

References

[1] Rishabh Balse, Viraj Kumar, Prajish Prasad, and Jayakrishnan Madathil Warriem. 2023. Evaluating the Quality of LLM-Generated Explanations for Logical Errors in CS1 Student Programs. In *16th ACM India Compute Conference* (<conf-loc>, <city>Hyderabad</city>, <country>India</country>, </conf-loc>). 49–54. doi:10.1145/3627217.3627233

[2] Wasihun Bezabih Bekele and Fikire Yohanes Ago. 2022. Sample size for interview in qualitative research in social sciences: A guide to novice researchers. *Research in Educational Policy and Management* 4, 1 (2022), 42–50.

[3] Philipp Brauner, Alexander Hick, Ralf Philipsen, and Martina Ziefle. 2023. What does the public think about artificial intelligence?—A criticality map to understand bias in the public perception of AI. *Frontiers in Computer Science* 5 (2023), 1113903.

[4] Nicole A Buzzetto-More. 2014. An examination of undergraduate student's perceptions and predilections of the use of YouTube in the teaching and learning process. *Interdisciplinary Journal of E-Learning and Learning Objects* 10, 1 (2014), 17–32.

[5] Donald Chinn, Judy Sheard, Angela Carbone, and Mikko-Jussi Laakso. 2010. Study habits of CS1 students: what do they do outside the classroom?. In *12th Australasian Conference on Computing Education*. Citeseer, 53–62.

[6] Pierpaolo Dondio and Suha Shaheen. 2019. Is stackoverflow an effective complement to gaining practical knowledge compared to traditional computer science learning?. In *11th International Conference on Education Technology and Computers*. 132–138.

[7] James Finnie-Ansley, Paul Denny, Andrew Luxton-Reilly, Eddie Antonio Santos, James Prather, and Brett A. Becker. 2023. My AI Wants to Know if This Will Be on the Exam: Testing OpenAI's Codex on CS2 Programming Exercises. In *25th Australasian Computing Education Conference* (Melbourne, VIC, Australia). 97–104. doi:10.1145/3576123.3576134

[8] Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How many interviews are enough? An experiment with data saturation and variability. *Field methods* 18, 1 (2006), 59–82.

[9] Philipp Haindl and Gerald Weinberger. 2024. Students' Experiences of Using ChatGPT in an Undergraduate Programming Course. *IEEE Access* 12 (2024), 43519–43529. doi:10.1109/ACCESS.2024.3380909

[10] Monique Hennink and Bonnie N Kaiser. 2022. Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social science & medicine* 292 (2022), 114523.

[11] Monique M Hennink, Bonnie N Kaiser, and Vincent C Marconi. 2017. Code saturation versus meaning saturation: how many interviews are enough? *Qualitative health research* 27, 4 (2017), 591–608.

[12] Muntasir Hoq, Yang Shi, Juho Leinonen, Damilola Babalola, Collin Lynch, Thomas Price, and Bita Akram. 2024. Detecting ChatGPT-Generated Code Submissions in a CS1 Course Using Machine Learning Models. In *55th ACM Technical Symposium on Computer Science Education*. 526–532. doi:10.1145/3626252.3630826

[13] 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 *26th Australasian Computing Education Conference* (Sydney, NSW, Australia). 39–48. doi:10.1145/3636243.3636248

[14] Ziheng Huang, Sebastian Gutierrez, Hemanth Kamana, and Stephen Macneil. 2023. Memory Sandbox: Transparent and Interactive Memory Management for Conversational Agents. In *36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA). Article 97, 3 pages. doi:10.1145/3586182.3615796

[15] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[16] Breanna Jury, Angela Lorusso, Juho Leinonen, Paul Denny, and Andrew Luxton-Reilly. 2024. Evaluating LLM-generated Worked Examples in an Introductory Programming Course. In *26th Australasian Computing Education Conference*. 77–86. doi:10.1145/3636243.3636252

[17] Stuart A Karabenick. 2003. Seeking help in large college classes: A person-centered approach. *Contemporary Educational Psychology* 28, 1 (2003), 37–58.

[18] Stuart A Karabenick and Jean-Louis Berger. 2013. Help seeking as a self-regulated learning strategy. *Applications of self-regulated learning across diverse disciplines: A tribute to Barry J. Zimmerman* (2013), 237–261.

[19] Stuart A Karabenick and Richard Stuart Newman. 2006. *Help seeking in academic settings: Goals, groups, and contexts*. Psychology Press.

[20] Hieke Keuning, Isaac Alpizar-Chacon, Ioanna Lykourentzou, Lauren Beehler, Christian Köppe, Imke de Jong, and Sergey Sosnovsky. 2024. Students' Perceptions and Use of Generative AI Tools for Programming Across Different Computing Courses. In *24th Koli Calling International Conference on Computing Education Research*. Article 14, 12 pages. doi:10.1145/3699538.3699546

[21] Anton J Kuzel. 1992. Sampling in qualitative inquiry. (1992).

[22] Sam Lau and Philip Guo. 2023. From" Ban it till we understand it" to" Resistance is futile": How university programming instructors plan to adapt as more students use AI code generation and explanation tools such as ChatGPT and GitHub Copilot. In *2023 ACM Conference on International Computing Education Research*. 106–121.

[23] Juho Leinonen, Paul Denny, Stephen MacNeil, Sami Sarsa, Seth Bernstein, Joanne Kim, Andrew Tran, and Arto Hellas. 2023. Comparing Code Explanations Created by Students and Large Language Models. In *2023 Conference on Innovation and Technology in Computer Science Education V. 1* (<conf-loc>, <city>Turku</city>, <country>Finland</country>, </conf-loc>) *(ITiCSE 2023)*. 124–130. doi:10.1145/3587102.3588785

[24] 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 *55th ACM Technical Symposium on Computer Science Education V. 1* (Portland, OR, USA) *(SIGCSE 2024)*. 750–756. doi:10.1145/3626252.3630938

[25] Andrew Luxton-Reilly. 2016. Learning to Program is Easy. In *2016 ACM Conference on Innovation and Technology in Computer Science Education* (Arequipa, Peru). 284–289. doi:10.1145/2899415.2899432

[26] Kara A Makara and Stuart A Karabenick. 2013. Characterizing college students' help seeking during computer-mediated communication. *Journal of Educational Psychology* 105, 2 (2013), 394–403.

[27] Eugine Tafadzwa Maziriri, Parson Gapa, and Tinashe Chuchu. 2020. Student perceptions towards the use of YouTube as an educational tool for learning and tutorials. *International Journal of Instruction* 13, 2 (2020), 119–138.

[28] Diba Mirza, Phillip T. Conrad, Christian Lloyd, Ziad Matni, and Arthur Gatin. 2019. Undergraduate Teaching Assistants in Computer Science: A Systematic Literature Review. In *2019 ACM Conference on International Computing Education Research* (Toronto ON, Canada). 31–40. doi:10.1145/3291279.3339422

[29] Sharon Nelson-Le Gall. 1981. Help-seeking: An understudied problem-solving skill in children. *Developmental review* 1 (1981), 224–246.

[30] Richard S Newman. 1994. Adaptive help seeking: A strategy of self-regulated learning. In *Self-regulation of learning and performance: Issues and educational applications*, Dale H Schunk and Barry J Zimmerman (Eds.). Lawrence Erlbaum Associates, 283–301.

[31] Jacob Penney, Pawan Acharya, Peter Hilbert, Priyanka Parekh, Anita Sarma, Igor Steinmacher, and Marco Aurelio Gerosa. 2025. Outcomes, Perceptions, and Interaction Strategies of Novice Programmers Studying with ChatGPT. In *7th ACM Conference on Conversational User Interfaces*. 1–15.

[32] Jacob Penney, João Felipe Pimentel, Igor Steinmacher, and Marco A Gerosa. 2023. Anticipating User Needs: Insights from Design Fiction on Conversational Agents for Computational Thinking. In *International Workshop on Chatbot Research and Design*. Springer, 204–219.

[33] Tung Phung, Victor-Alexandru Pădurean, Jos é Cambronero, Sumit Gulwani, Tobias Kohn, Rupak Majumdar, Adish Singla, and Gustavo Soares. 2023. Generative AI for Programming Education: Benchmarking ChatGPT, GPT-4, and Human Tutors. In *2023 ACM Conference on International Computing Education Research - Volume 2* (Chicago, IL, USA). 41–42. doi:10.1145/3568812.3603476

[34] James Prather, Juho Leinonen, Natalie Kiesler, Jamie Gorson Benario, Sam Lau, Stephen MacNeil, Narges Norouzi, Simone Opel, Vee Pettit, Leo Porter, et al. 2024. Beyond the Hype: A Comprehensive Review of Current Trends in Generative AI Research, Teaching Practices, and Tools. *arXiv preprint arXiv:2412.14732* (2024).

[35] Thomas W. Price, Zhongxiu Liu, Veronica Cateté, and Tiffany Barnes. 2017. Factors Influencing Students' Help-Seeking Behavior while Programming with Human and Computer Tutors. In *2017 ACM Conference on International Computing Education Research*. 127–135. doi:10.1145/3105726.3106179

[36] Allison M Ryan and Paul R Pintrich. 1997. Should I ask for help? The role of motivation and attitudes in adolescents' help seeking in math class. *Journal of Educational Psychology* 89, 2 (1997), 329–341.

[37] Judy Sheard, Paul Denny, Arto Hellas, Juho Leinonen, Lauri Malmi, and Simon. 2024. Instructor Perceptions of AI Code Generation Tools - A Multi-Institutional Interview Study. In *55th ACM Technical Symposium on Computer Science Education V. 1* (Portland, OR, USA). 1223–1229. doi:10.1145/3626252.3630880

[38] James Skripchuk, John Bacher, and Thomas Price. 2024. An Investigation of the Drivers of Novice Programmers' Intentions to Use Web Search and GenAI. In *2024 ACM Conference on International Computing Education Research* (Melbourne, VIC, Australia). 487–501. doi:10.1145/3632620.3671112

[39] Francesco Walker, Matteo Favetta, Linde Hasker, and Richard Walker. 2024. They Prefer Humans! Experimental Measurement of Student Trust in ChatGPT. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA). Article 325, 7 pages. doi:10.1145/3613905.3650955

[40] Jeannette M. Wing. 2006. Computational thinking. *Commun. ACM* 49, 3 (mar 2006), 33–35. doi:10.1145/1118178.1118215

[41] David Wong-Aitken, Diana Cukierman, and Parmit K. Chilana. 2022. "It Depends on Whether or Not I'm Lucky" How Students in an Introductory Programming Course Discover, Select, and Assess the Utility of Web-Based Resources. In *27th ACM Conference on on Innovation and Technology in Computer Science Education* (Dublin, Ireland). 512–518. doi:10.1145/3502718.3524751

[42] Cynthia Zastudil, Magdalena Rogalska, Christine Kapp, Jennifer Vaughn, and Stephen MacNeil. 2023. Generative AI in Computing Education: Perspectives of Students and Instructors. In *2023 IEEE Frontiers in Education Conference (FIE)*. 1–9. doi:10.1109/FIE58773.2023.10343467