An Experience Report on a Pedagogically Controlled, Curriculum-Constrained AI Tutor for SE Education

Lucia Happe
lucia.happe@kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

Dominik Fuchß
dominik.fuchss@kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

Luca Hüttner
luca.huettner@student.kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

Kai Marquardt
kai.marquardt@kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

Anne Koziolek
koziolek@kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

Abstract

The integration of artificial intelligence (AI) into education continues to evoke both promise and skepticism. While past waves of technological optimism often fell short, recent advances in large language models (LLMs) have revived the vision of scalable, individualized tutoring. This paper presents the design and pilot evaluation of *RockStartIT Tutor*, an AI-powered assistant developed for a digital programming and computational thinking course within the RockStartIT initiative. Powered by GPT-4 via OpenAI's Assistant API, the tutor employs a novel prompting strategy and a modular, semantically tagged knowledge base to deliver context-aware, personalized, and curriculum-constrained support for secondary school students.

We evaluated the system using the Technology Acceptance Model (TAM) with 13 students and teachers. Learners appreciated the low-stakes environment that encouraged them to ask questions and receive scaffolded guidance. Educators emphasized the system's potential to reduce cognitive load during independent tasks and complement classroom teaching. Key challenges include prototype limitations, a small sample size, and the need for long-term studies with the target age group.

Our findings highlight a pragmatic approach to AI integration that requires no model training — using structure and prompts to shape behavior. We position AI tutors not as teacher replacements but as enabling tools that extend feedback access, foster inquiry, and support what schools do best: help students learn.

CCS Concepts

• **Applied computing** → **Education**; • **Social and professional topics** → **Computer science education**; **Computer engineering education**; **Software engineering education**; **Computing education**.

Keywords

AI tutors, GPT-4, chatbots, large language models, LLMs, software engineering education, se education, computer science education, technology acceptance model, digital learning

ACM Reference Format:

Lucia Happe, Dominik Fuchß, Luca Hüttner, Kai Marquardt, and Anne Koziolek. 2026. An Experience Report on a Pedagogically Controlled, Curriculum-Constrained AI Tutor for SE Education. In *2026 IEEE/ACM 48th International Conference on Software Engineering (ICSE-SEET '26), April 12–18, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3786580.3786947

1 Introduction

In 1913, Thomas Edison famously predicted that motion pictures would revolutionize education and render books in schools obsolete [25]. A century later, similar optimism surrounds artificial intelligence (AI) in education. Modern AI-powered chatbots have revived long-standing visions of a personal tutor for every student — capable of answering questions, guiding inquiry, and personalizing instruction at scale. Yet, as with past technological interventions, the transformative promise of AI often falters in practice — due to limited evaluation, poor classroom integration, or misalignment with pedagogical goals [5, 22].

Despite these challenges, AI retains significant potential, especially when applied thoughtfully to support rather than replace instruction. In contexts where teacher attention is limited and learning is increasingly digital and asynchronous, AI tutors can fill critical gaps. This is particularly relevant in software engineering (SE) education, where students frequently work independently and face complex problem-solving tasks without immediate support [9]. Initiatives like RockStartIT [12] show that engagement in SE increases when computing is taught through real-world, interdisciplinary problems. However, maintaining that engagement requires more than content delivery — it demands responsive guidance.

AI tutors offer scalable, individualized support that is always available, nonjudgmental, and capable of providing immediate feedback [26]. Such systems may lower psychological barriers and encourage low-stakes, exploratory learning [15]. Especially in SE education — where self-study is the norm — AI tutors can be a valuable alternative to learning in isolation.

ICSE-SEET '26, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2423-7/2026/04
https://doi.org/10.1145/3786580.3786947

Still, as Reich notes, "talking to robots is boring" [22]. Learning thrives in social, human-centered environments. AI tutors cannot replace collaborative classroom discourse, peer interaction, or teacher mentorship. However, in self-paced SE environments, they can offer a meaningful improvement over silence. Moreover, dialogue with AI can foster metacognitive skills — such as articulating reasoning, formulating questions, and iteratively refining ideas — that are essential to SE practice [29].

Rather than proposing AI tutors as replacements for educators, this work embraces a complementary vision: AI as an enabler that supports learner autonomy, amplifies instructional reach, and allows teachers to focus on what they do best — facilitating insight, feedback, and human connection [20].

This paper presents the design and evaluation of a pedagogically controlled AI tutor developed for a RockStartIT course on programming and computational thinking. The tutor uses OpenAI's GPT-4 model via the Assistant API, embedded in a device-independent web interface. Instead of relying on model fine-tuning or external search infrastructure, the system combines a modular, semantically tagged knowledge base with carefully engineered prompts to control behavior, guide dialogue, and restrict output to approved instructional content. This approach enables safe, personalized, and curriculum-aligned tutoring without requiring any custom training or internal model modifications.

Research Goal: Our goal is not to replace educators but to offer a **lightweight and pragmatic architecture** that empowers teachers and learners alike. The AI tutor should serve as an enabler: **filling gaps in individual support, reducing learner hesitation, and communicating in an age-appropriate, motivating, and accessible language**. Thus, it enables real-time interaction while **avoiding hallucinations**, **enforcing curriculum boundaries**, and **promoting inquiry-driven learning** — particularly in asynchronous and independent study contexts common in SE education. We evaluate the system through the lens of the Technology Acceptance Model (TAM) [7], combining quantitative and qualitative feedback from both students and teachers. Our findings inform short-term design improvements and point toward future work in adaptive scaffolding, trust calibration, pedagogical safety, and broader curricular integration.

2 Background and Related Work

In this section, we provide an overview of the relevant background and related work that informs our research on AI tutors in secondary software engineering education.

2.1 AI Assistants for Education

Large language models (LLMs) such as GPT-3.5 and GPT-4 have gained traction as tools for supporting programming and computational thinking education. AI-powered assistants enable natural language interaction and can serve as virtual tutors, providing timely, context-sensitive feedback across a wide range of tasks.

Dai et al. [6] examined ChatGPT for automated student feedback, assessing readability, instructional similarity, and learning impact. They found that ChatGPT produced clear, coherent feedback aligned with instructor evaluations and offered useful insights into student thinking.

Similarly, Frankford et al. [8] developed an AI tutor based on GPT-3.5-Turbo to support students in software engineering exercises, aiming to reduce the workload on teaching assistants. Integrated into the APAS Artemis system, the AI tutor offered faster response times and 24/7 availability. However, students' perceptions of helpfulness varied, and concerns were raised regarding generic responses and their potential impact on deep learning.

Bassner et al. [3] developed Iris, an AI tutor using chain-of-thought prompting with access to student code. User studies showed it clarified concepts and provided hints without revealing answers.

Hellas et al. [13] examined the use of GPT-3.5 and Codex for code feedback. While the models detected at least one genuine issue in 90% of cases, they also produced false positives and often failed to identify all problems. Instructors frequently needed to refine AI feedback before it could be useful to students, highlighting the limitations of LLMs in their current state.

In a broader educational context, Kestin et al. [16] conducted a large-scale randomized controlled trial comparing an AI tutor with active learning in physics education. The AI-supported group showed significantly greater learning gains and reported higher engagement, emphasizing the potential of well-designed AI tutors to complement traditional instruction.

While promising, most prior studies have focused on higher education and technical domains. There remains a notable gap in research on AI tutors for younger learners (aged 13–16), especially in real-world deployments where cognitive development, trust, and instructional needs differ significantly.

2.2 Pedagogical Alignment and Instructional Control

Despite rapid progress, many current AI tutoring systems rely on generic prompts and unstructured content, lacking alignment with educational goals. Few systems offer mechanisms to control instructional behavior, adapt scaffolding levels, or ensure that the AI adheres to curriculum boundaries. Our work addresses these limitations by embedding domain-specific instructions and using a modular, semantically tagged knowledge base to guide AI behavior.

To our knowledge, no previous study has deployed an AI tutor that uses structured tagging and traversal logic to dynamically adapt behavior while ensuring fidelity to pre-approved instructional content. This allows for fine-grained pedagogical control without requiring custom model training or retraining.

2.3 Human-AI Collaboration and Trust Calibration

Recent research has begun to explore how AI tutors can be integrated into classroom dynamics without displacing human educators. Holstein and et al. [15] emphasize the importance of co-designing AI systems that empower teachers and support their instructional strategies. Similarly, trust calibration has emerged as a concern in AI-supported learning [21]. Learners may over- or under-trust AI, especially when system transparency is low.

Our implementation supports teacher-aligned control through behavior tagging, content restriction, and transparent response sourcing. We also collect open-ended feedback on students' and teachers' perceptions of trustworthiness and limitations.

2.4 RockStartIT Interdisciplinary SE Courses

RockStartIT is an initiative that provides engaging and accessible online SE courses for school-aged learners. Based on the IDEA (Interdisciplinary, Diverse, Exploratory, Active) framework [10, 11], they combine real-world challenges with computing instruction.

RockStartIT expeditions immerse students in interdisciplinary journeys that encourage inquiry, teamwork, and experimentation. In the "Save the Bees" expedition, students investigate colony collapse disorder through web development, data analysis, and AI. They learn programming and problem-solving by working on tasks such as creating awareness websites, analyzing bee flight data, and evaluating hive health using SQL and image data. These explorations are designed to connect CS skills with meaningful contexts.

2.5 Technology Acceptance in Educational Contexts

The Technology Acceptance Model (TAM) [7] is a framework for understanding how users adopt educational technologies. It identifies Perceived Usefulness (PU), Perceived Ease of Use (PEOU), and Intention to Use (IU) as central factors influencing acceptance.

While TAM has been widely applied to chatbots and educational tools, most studies rely on simulated interactions or prototype systems. Few evaluate fully operational systems in authentic classroom contexts. This gap underscores the need for empirical studies that capture the real-world challenges and contextual influences on AI acceptance.

2.6 Positioning This Work

Our research advances the field in several ways:

- We deploy a fully functional AI tutor using a modular and behavior-tagged knowledge base — without retraining or external search.
- We evaluate the system in a real-world SE course with secondary students and teachers.
- We extend TAM-based evaluation to include open-ended perceptions of trust, adaptability, and instructional fit.

By focusing on grounded deployment and educational control, our work contributes practical insights into how AI tutors can be safely and effectively integrated into school settings.

3 System Architecture and Implementation

Rather than building a new AI model, the work presents a novel, lightweight, and powerful framework for turning general-purpose LLMs into pedagogically effective tutors using only clever architecture design, tool restriction, and instructional prompting. This makes it highly replicable and safe for educational use. The AI tutor was designed to provide pedagogically aligned, adaptive, and context-sensitive support for learners in a modular and scalable manner. This section outlines the system architecture, key components, and the core innovations behind its implementation.

3.1 Design Goals and Requirements

The AI tutor was designed to support students aged 14–16 participating in the RockStartIT course. The key design requirements were as follows:

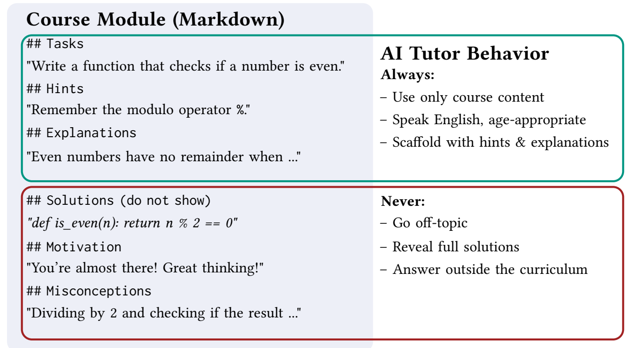

Figure 1: Simplified example of tagged Markdown structure for a learning content.

- Provide context-aware, personalized support **strictly grounded in the course content**, and with mechanisms to **prevent breakout into unrelated topics or general knowledge**.
- Communicate in an age-appropriate, motivating, and accessible language suitable for lower secondary school students.
- Promote **inquiry and independent thinking by offering hints and guidance instead of direct solutions**.
- Ensure seamless access across devices without requiring student login or registration.
- Enable real-time interaction while **avoiding hallucinations and enforcing curriculum boundaries**, ensuring pedagogical safety and content fidelity.

3.2 Overall Architecture

The AI Tutor was implemented as a modular, web-based educational assistant powered by OpenAI's GPT-4o via the Assistant API. It follows a client-server model with three main components:

- **Frontend:** A device-independent chat interface built with React and Next.js, using the Chat UI Kit React library. It allows students to intuitively submit questions and displays real-time streamed responses from the AI.
- **Backend:** Next.js API routes that securely manage API keys, user sessions, conversation threads, and message routing.
- **AI Layer:** Utilizes OpenAI's GPT-4o via the Assistant API. It is configured with pedagogical instructions and connected to a structured knowledgebase composed of Markdown-formatted course content uploaded as file tools.

3.3 Structured Knowledgebase Integration

A core innovation of the KI Tutor lies in its use of OpenAI's Assistant API tool-calling mechanism to strictly constrain responses to a structured, curriculum-aligned knowledgebase. This eliminates hallucinations and ensures pedagogical safety. The knowledgebase is composed of modular, Markdown-formatted files that encapsulate the RockStart IT course content (as illustrated in Figure 1), including concepts, scaffolded hints, learning tasks, and motivational prompts. These files are uploaded as a dedicated `file_search` tool accessible to the Assistant API, ensuring that all responses remain grounded in predefined instructional material.

Tagged Modular Format: The course content was organized into multiple Markdown files, each corresponding to a specific learning module. Within each file, content was segmented into pedagogically meaningful units using consistent second-level Markdown headers such as `tasks`, `hints`, `explanations`, `solutions` (explicitly marked as "do not show"), `motivation`, and `misconceptions`. Please be aware that in our original study all terms were in German.

This structure served several key purposes: (i) it enabled modular, curriculum-aligned retrieval through consistent file and section formatting; (ii) it ensured instructional control by mapping each content type to defined tutor behaviors; and (iii) it supported scalable expansion of the knowledge base without requiring changes to the underlying system logic. The assistant used a layered traversal strategy: beginning with hints and explanations, escalating only when needed, and withholding direct solutions unless explicitly permitted. Motivational feedback was delivered upon completion of conceptual milestones. This scaffolding approach reflects principles from intelligent tutoring systems and supports adaptive learning paths. Unlike general RAG or fine-tuning pipelines, this implementation offers tight control over knowledge access and response generation—ensuring factual precision, pedagogical alignment, and safety without training a custom model.

3.4 Instruction-Based Pedagogical Behavior

To align the AI tutor's behavior with educational goals, the system employs a carefully crafted instruction set using OpenAI's system prompt mechanism. Key instructions include:

- Respond only in German using age-appropriate language.
- Limit responses strictly to the provided course content.
- Avoid giving direct solutions, especially for tagged answers.
- Support learners with Socratic questioning and scaffolded hints.
- Maintain a motivating and friendly tone (students aged 14–16).

Prompt Engineering and Behavior Control: Hallucination, verbosity, and over-answering were reduced through iterative prompt tuning; the final system prompt (Prompt 1) adds explicit guidance.

To align AI responses with pedagogical intent, the system incorporates semantic behavior tags (e.g., `task`, `explanation`, `solution`) within the knowledge base. These tags inform the assistant how to respond based on the instructional function of the content. When a student submits a query, the system:

(1) Classifies the user's intent (e.g., clarification, task assistance, conceptual exploration).
(2) Filters and retrieves content segments with matching tags.
(3) Constructs a response using layered instructional strategies tailored to the content type and learner context.

For example, task-related queries trigger a *hint-first* scaffolding strategy, while conceptual questions follow an *explanation-then-analogy* pattern to deepen understanding. This combination of content-based restrictions and tag-driven instructional behavior — achieved entirely through prompt engineering — offers a lightweight yet powerful method for developing safe, adaptive, and curriculum-aligned AI tutors without requiring model fine-tuning.

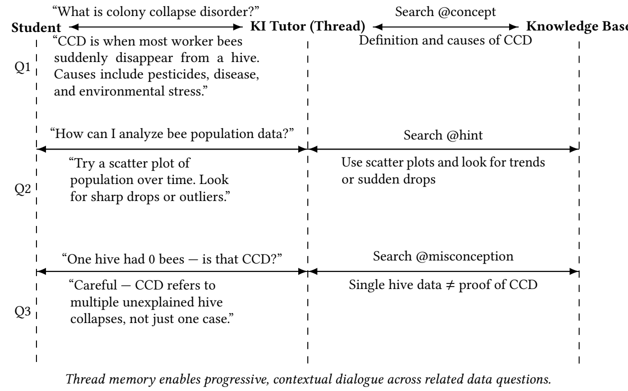

Figure 2: Tutor interaction for RockStartIT's bee data science module — guiding from concept to analysis and clarification.

Prompt 1: System Prompt of AI Tutor

This GPT is a chatbot tutor named AI Tutor, designed to assist students with questions related to an online course called "Rettet die Bienen!".

It should provide clear, concise answers, offer additional explanations when necessary, and be patient and encouraging. It should ALWAYS and only use the Information in the Knowledgebase to answer the questions. Only answer questions relevant to the course.

All responses should be written in German and kept concise. The students are around the age of 14 to 16 years old, so answer in that niveau.

When asked for solutions to a question that is in the "Antworten" part of the files in the knowledge base, AI Tutor should NEVER give out the solution directly but instead guide and motivate students to find the answers themselves by giving tips, asking guiding questions, and encouraging critical thinking without directly providing the answers.

3.5 Context Preservation and Adaptive Interaction

Using OpenAI's thread model, each user session maintains conversational memory across exchanges (see Figure 2). This enables the tutor to refer to prior context for coherence and continuity, tailor scaffolding based on learner progress, and sustain meaningful multi-turn educational dialogue.

The system applies an adaptive content assembly pipeline:

(1) **Intent Matching:** Detects learner need (e.g., help, clarification).
(2) **Tag-Based Retrieval:** Locates pedagogically relevant snippets.
(3) **Response Construction:** Builds scaffolded responses that are adapted to prior input.

This enables personalized instructional paths and controlled pedagogical responses. Thread handling via Assistant API enables a "conversational memory" without persistent user tracking.

3.6 Real-Time Response Streaming

To enhance engagement and reduce latency, the system uses OpenAI's streaming API. Rather than waiting for full responses, output is transmitted token-by-token, simulating natural dialogue pacing:

- **Reduces latency:** Feedback appears instantly, improving flow.
- **Simulates live interaction:** Output feels human and reactive.
- **Boosts attention:** Unfolding answers promote continued focus.

This reinforces the tutor's responsiveness and real-time support behavior. This design choice makes the AI feel more human and responsive compared to static full-message replies.

3.7 Novelty and Strengths of the Approach

The AI Tutor introduces a lightweight yet powerful blueprint for integrating LLMs into educational contexts. Its core novelty lies in harmonizing the flexibility of generative AI with structured pedagogical control — achieving both adaptability and instructional safety without requiring model fine-tuning.

Introduced improvements include:

- **Curriculum-Constrained Output:** Responses are grounded in a predefined, structured knowledge base, eliminating hallucinations and ensuring factual and curricular alignment.
- **Instructionally Tuned Prompts:** Behavior is explicitly shaped through domain-specific system instructions, guiding the AI to scaffold learning, avoid solution-giving, and engage students in reflection.
- **Context-Aware Dialogue:** Thread-based memory supports coherent, multi-turn conversations that adapt to learner progress and input history.
- **Live Response Streaming:** Token-level response streaming creates a dynamic, conversational experience, enhancing student engagement and reducing perceived latency.

Architectural strengths include:

- **Transparency and Traceability:** A modular, tagged knowledge base allows educators to inspect and adjust content-response mappings, fostering explainability and trust.
- **Scalable Design:** New topics or modules can be integrated by adding tagged content — without modifying traversal logic or retraining the model.
- **Pedagogical Controllability:** Traversal strategies and semantic tags act as interpretable levers to steer the AI toward didactically sound behaviors.
- **Contextual Personalization:** The tutor dynamically assembles responses based on query intent, learner history, and content tags — enabling fine-grained instructional adaptation.

Using structured content control, behavior-aware prompting, and contextual thread management, the AI Tutor delivers an interpretable, extensible, and pedagogically robust tutoring experience. The approach is applicable beyond a single course or age group, offering a reusable framework for safe and effective AI integration in education.

Table 1: Participant Overview

Participants	Teachers	Learners	Total
Total	6	7	13
Male	5	5	10
Female	1	2	3
Average age	32.5	23.7	27.8

4 Evaluation

To assess the functionality and usefulness of the AI tutor, we conducted an empirical evaluation with students and teachers, using the tutor as a study companion during a RockStartIT course. The aim of this study is to evaluate the acceptance and use of the AI tutor by teachers and learners, using the Technology Acceptance Model (TAM) proposed by Davis [7]. Special attention is given to how the AI tutor is perceived, the factors that influence its acceptance, and the extent to which it supports the learning process.

4.1 Methodology

We employed a mixed-methods approach, collecting quantitative feedback through a TAM-inspired questionnaire and gathering qualitative data via four open-ended questions. Quantitative items, 12 (plus four additional for teachers) Likert-scale items, measured three core constructs:

- **Perceived Usefulness (PU)** – e.g., "The tutor helped me understand the topic."
- **Perceived Ease of Use (PEOU)** – e.g., "It was easy to ask the tutor questions."
- **Intention to Use (IU)** – e.g., "I would like to use a similar tutor in the future."

Each item was rated on a 5-point scale from 1 (strongly disagree) to 5 (strongly agree). Teachers and learners completed the questionnaire after using the AI tutor in class. The AI tutor was used for 30–45 minutes on participants' own devices. They were encouraged to explore freely, ask any questions, and follow their curiosity. Afterward, they completed a structured questionnaire.

The open-ended questions explored challenges, suggestions for improvement, helpful/unhelpful scenarios, and attitudes toward the tutor's future role in education.

4.2 Participants

A total of 13 participants took part in the study, including six (prospective) teachers with experience in computer science education and seven learners (see Table 1), most of whom had no prior programming experience. Teachers were either course instructors or external computer science educators familiar with the curriculum. Among the teachers, five were male and one female; among learners, five were male and two female. The average age was 32.5 years for teachers and 23.7 years for learners.

Demographic Questions:

(1) Are you currently working as a teacher or in teacher training?
(2) What gender do you identify with?
(3) How old are you?

Table 2: Selected Questionnaire Results

Item	Average	SD
PU8: The AI tutor can be a useful addition to my teaching materials.	4.5	0.54
PEOU5: I find the AI tutor easy to use.	4.6	0.50

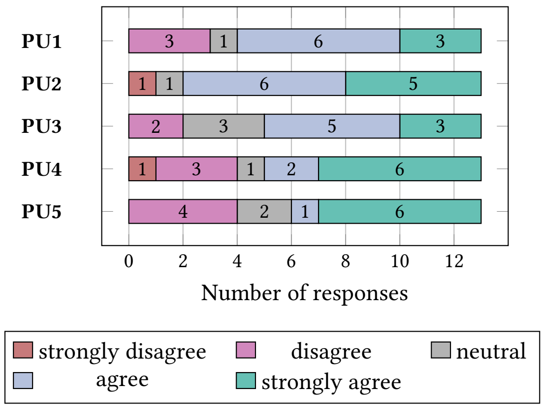

Figure 3: Perceived Usefulness ratings from teachers/learners

These questions were used to analyze responses by demographic groups. The role (teacher vs. learner) determined which follow-up questions were shown.

4.3 Quantitative Results

Overall, the AI tutor was rated as easy to use and useful. Students especially appreciated the low threshold for asking questions. Table 2 shows a summary of the selected questionnaire results.

The responses indicated strong acceptance among learners, with most items scoring above 4. Teachers also rated the tutor positively, noting its potential as a support tool rather than a replacement for instruction.

4.3.1 Perceived Usefulness. This subsection analyzes participant responses related to *Perceived Usefulness (PU)*. Figure 3 shows the distribution of responses across all PU items, and Table 3 summarizes the corresponding descriptive statistics.

Overall, participants rated the AI tutor as useful, particularly in terms of its reliability and potential to support learners. The statement PU3 received the highest level of agreement, followed closely by PU1 and PU2.

Notably, PU1 and PU4 showed a strong positive correlation ($r = 0.77$), while PU2 and PU4 demonstrated a moderate correlation ($r = 0.58$). These results suggest that participants who found the AI tutor reliable and supportive were also more likely to see it as a meaningful aid in completing learning tasks.

Teacher responses were particularly favorable: items PU2 and PU8 received unanimous agreement, highlighting the system's perceived consistency and relevance to classroom practice. However, the statement PU6 received more mixed responses—two out of six teachers expressed slight skepticism. This suggests that while teachers acknowledge the tutor's instructional value, its integration into daily teaching workflows remains an open question.

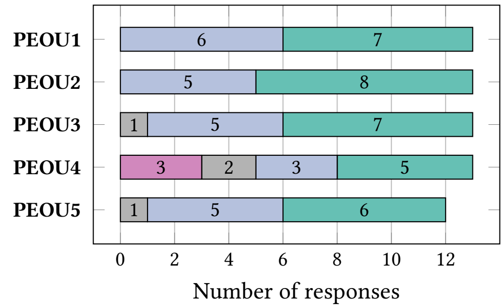

Figure 4: Perceived Ease of Use (PEOU) ratings

Learner ratings, though generally positive, displayed greater variance. PU1 and PU2 were the most highly rated, while PU4 and PU5 received more moderate evaluations (mean $M = 3.43$), still trending toward agreement. This variation likely reflects the mismatch between the tutor's intended audience (grades 7–10) and the higher-level academic background of the surveyed university students. Some learners noted that the tutor's responses were too generic or overly cautious—an intentional design choice aimed at promoting reflection rather than simply delivering answers.

In summary, the AI tutor was perceived as useful, particularly in supporting understanding and providing consistent feedback. Teachers in particular valued its reliability and potential to complement instruction, though questions remain about its direct utility in streamlining their workload. Learners appreciated its clarity and guidance, though the design trade-off between answer provision and Socratic scaffolding led to mixed reactions. These findings underline the importance of aligning AI tutor design with both pedagogical goals and the specific needs of target user groups.

4.3.2 Perceived Ease of Use. Participants generally rated the AI tutor as easy to use (see Figure 4). The most positively rated items were PEOU1 and PEOU2. The only item to receive notable disagreement was PEOU4, which reflects a degree of variability in participants' perceptions of system responsiveness.

Learners tended to rate ease of use slightly more favorably than teachers, though both groups expressed consistent agreement with most statements. Among learners, PEOU1 received the strongest endorsement, with six participants selecting the highest level of agreement. Among teachers, all items except PEOU4 were rated positively, while PEOU4 drew two slightly negative responses, indicating occasional mismatches between user intent and system behavior.

Overall, the AI tutor was perceived as intuitive and accessible. Participants found the chat-based interface familiar and easy to navigate — especially students, who may be more accustomed to conversational digital tools. Teachers also responded positively, although their slightly lower ratings on PEOU4 suggest that occasional functional limitations may have interfered with expectations of seamless interaction.

Table 3: Descriptive statistics for Perceived Usefulness (PU), Perceived Ease of Use (PEOU), and Intention to Use (IU) items

Item	N	Mean	Median	SD
PU1: The AI tutor helped me understand the topics more easily.	13	3.69	4	1.11
Teachers only	6	3.67	4	1.03
Learners only	7	3.71	4	1.25
PU2: The AI tutor responded reliably to my inputs.	13	4.08	4	1.12
Teachers only	6	4.50	4.5	0.55
Learners only	7	3.71	4	1.38
PU3: The AI tutor improved the learning experience.	13	3.69	4	1.03
Teachers only	6	3.83	4	0.75
Learners only	7	3.57	4	1.27
PU4: I think the AI tutor can meaningfully support learners with tasks.	13	3.69	4	1.49
Teachers only	6	4.00	4.5	1.27
Learners only	7	3.43	4	1.71
PU5: I find the AI tutor useful.	13	3.69	4	1.38
Teachers only	6	4.00	4.5	1.27
Learners only	7	3.43	3	1.51
PU6: I believe an AI tutor like this can be useful in my work/teaching. *(teachers only)*	6	3.50	4	1.23
PU7: Using such an AI tutor in class makes my job easier. *(teachers only)*	6	3.67	3.5	1.21
PU8: The AI tutor can be a useful addition to my teaching materials. *(teachers only)*	6	4.50	4.5	0.55
PEOU1: Interacting with the AI tutor is clear and understandable.	13	4.54	5.0	0.52
Teachers only	6	4.12	4.0	0.41
Learners only	7	4.86	5.0	0.38
PEOU2: I find the AI tutor easy to use overall.	13	4.62	5.0	0.51
Teachers only	6	4.5	4.5	0.55
Learners only	7	4.71	5	0.49
PEOU3: It is easy to use the AI tutor even without technical knowledge.	13	4.46	5.0	0.66
Teachers only	6	4.5	4.5	0.55
Learners only	7	4.43	5	0.79
PEOU4: The AI tutor does what I want without issues.	13	3.77	4.0	1.24
Teachers only	6	3.67	4.0	1.37
Learners only	7	3.86	4	1.22
PEOU5: I find the AI tutor easy to use.	12	4.42	4.5	0.67
Teachers only	5	4.6	5.0	0.55
Learners only	7	4.29	4	0.76
IU1: Assuming I had access to such an AI tutor, I would intend to use it.	13	3.85	4.0	1.41
Teachers only	6	3.83	4.0	1.27
Learners only	7	3.86	5	1.68
IU2: If I had access to this AI tutor, I predict I would use it.	13	3.62	4.0	1.26
Teachers only	6	3.5	3.5	1.38
Learners only	7	3.71	4	1.25
IU3: I can imagine using such an AI tutor regularly in my work/classroom. *(teachers only)*	6	3.83	4.0	1.17

These results support the design decision to use a conversational interface. The generally high ratings across both groups indicate that the user interface posed minimal barriers to adoption. The system's clarity and straightforward interaction style contributed to a strong perception of ease of use—an essential factor for engagement, particularly in educational environments where user confidence and comfort directly impact learning outcomes.

4.3.3 Intention to Use. Figure 5 summarizes participants' responses to items related to *Intention to Use (IU)*. Both IU1 and IU2 were rated positively by the majority of participants, indicating a general openness to future use. IU1 saw slightly higher overall agreement, but also the only strongly negative rating in this category.

Among learners, responses were slightly more variable, reflecting differing levels of enthusiasm. In contrast, teachers provided more consistently positive ratings. The teacher-specific item IU3 also received strong agreement, further reinforcing the system's perceived practical utility. A notable finding is the strong positive correlation ($r = 0.79$) between participants' perception of system reliability (PEOU4) and their stated intention to use the tutor in the future (IU2).

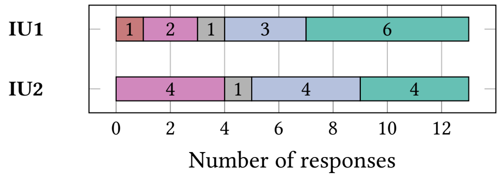

Figure 5: Intention to Use (IU) ratings from teachers / learners

This suggests that perceived consistency in system behavior plays a crucial role in shaping user trust and the likelihood of adoption. Additionally, among teachers, a moderate positive correlation ($r = 0.48$) emerged between age and IU3, indicating that older educators in the sample were slightly more inclined to envision regular classroom use — possibly reflecting a greater appreciation for tools that support workload management.

Overall, the intention to use the AI tutor was high across both groups. Students showed enthusiasm likely linked to their digital fluency and comfort with conversational interfaces. Teachers' responses were especially encouraging: despite their critical role as gatekeepers of classroom technology, most expressed a clear willingness to integrate the tutor into their instructional practice. Open-ended responses echoed these sentiments, highlighting the tutor's value for delivering one-on-one support in settings with limited teaching staff.

However, several participants also cautioned that the tutor's effectiveness depends heavily on pedagogical alignment. In rigidly structured, answer-driven classrooms, its open-ended guidance may offer limited benefit. This underscores the importance of flexible deployment strategies that empower teachers to tailor AI support to diverse instructional contexts and goals.

4.3.4 Qualitative Feedback. The following open-ended questions were designed to complement the TAM constructs and provide more in-depth insight into participant experiences.

Q1 **What challenges did you encounter while using the AI tutor?** Targets perceived ease of use (PEOU), exploring usability obstacles and user experiences.

Q2 **Do you have suggestions for how the AI tutor could be improved?** Gathers direct feedback for iterative development and adaptation to user needs.

Q3 **Were there situations where the AI tutor was especially helpful or less helpful?** Addresses perceived usefulness (PU) and helps identify strong or weak use cases.

Q4 **Do you think AI tutors like this could be a useful addition to future teaching? Why or why not?** Aims to understand long-term acceptance and perceived value (IU), especially from the perspective of educators.

Representative Quotes:

- *Q_1: "It is sometimes unclear what the tutor's level of knowledge is."*
- *Q_2: "Example questions that one can ask."*
- *Q_3: "Was very helpful in summarizing the entire content."*
- *Q_3: "Especially helpful: clarifying unknown terms through paraphrasing, obtaining additional information."*
- *Q_3: "It helped me understand why I answered questions incorrectly. That way I could trace back where my thinking error was."*
- *Q_4: "Yes, absolutely. Using the AI tutor is equivalent to one-on-one support, which is not possible through a teacher in regular classes. It relieves the teacher and enriches the students."*
- *Q_4: "...already asking the 'right' questions requires a deeper engagement with the content."*

The open-ended responses provided rich insights into participants' experiences, expectations, and suggestions for the AI tutor. Feedback revealed a strong interest in improving the system's usability, transparency, and instructional value—alongside an overall positive perception of its educational potential.

Usability and Onboarding. One of the most consistent themes was the need for improved onboarding. Both learners and teachers reported uncertainty regarding how to interact effectively with the tutor. Participants suggested the inclusion of example prompts, guided walkthroughs, and clearer indications of the tutor's capabilities and limitations. Teachers, in particular, emphasized the importance of managing expectations by clearly defining what the AI can and cannot do from the outset.

Context Awareness and Curriculum Integration. Several responses highlighted the tutor's limited understanding of the lesson context, which occasionally resulted in generic or repetitive answers. Participants emphasized that deeper curriculum integration—such as synchronizing with current chapters or providing task-specific scaffolding—would enhance the system's relevance and effectiveness. Teachers also expressed interest in features that would allow them to guide or constrain the tutor's focus in alignment with instructional goals.

Instructional Support and Modalities. Many users appreciated the tutor's ability to explain concepts, offer alternative perspectives, and clarify misunderstandings—especially for novice learners. However, there was a clear call for more example-based explanations and multimodal support, such as the ability to interpret images or present graphical explanations. Participants noted that such features would improve accessibility and support a broader range of learning styles. Teachers also valued the tutor's capacity to support learners with reading or writing difficulties, citing its robust handling of minor input errors.

Autonomy and Engagement. Teachers and students alike recognized the AI tutor's potential to foster independent learning and reduce the need for constant teacher supervision—particularly in large or mixed-ability classrooms. The system was perceived as emotionally safe, encouraging learners to ask questions freely without fear of judgment. Several comments noted that this availability and neutrality helped build confidence, especially among learners who were less confident.

Concerns and Cautions. Despite the generally positive reception, participants raised valid concerns. Some cautioned that the tutor might encourage surface-level engagement if learners relied too heavily on direct answers rather than developing problem-solving strategies. Others noted occasional repetitiveness or ambiguity in responses. Teachers were especially aware of the risk of overreliance. They advocated for a complementary integration model, where the tutor supports but does not replace traditional teaching and group-based learning.

Design Implications. Taken together, the feedback points to a range of short-term and long-term development priorities. In the near term, improvements should focus on onboarding, prompt design, and interface clarity. In the long term, participants called for deeper curricular integration, adaptive capabilities, and multimodal interaction (e.g., image input or voice response). These enhancements could significantly strengthen the tutor's educational impact and ensure its alignment with diverse teaching practices.

Overall, qualitative responses highlighted the strengths of a tutor in clarity, availability, inclusiveness, and encouragement of learner autonomy, providing actionable insights for refinement. The findings reinforce the potential of AI tutors to complement human instruction and underscore important design considerations for their effective adoption in educational settings.

5 Discussion

At the heart of this project lies a pragmatic philosophy: to design an AI tutor that complements rather than replaces human instruction, built not through costly fine-tuning but through prompt engineering and structured content curation. The result is a system that offers pedagogical guidance, respects curricular boundaries, and can be deployed in real classrooms today — not just in theory.

5.1 Perceived Value and Role of the Tutor

Despite a modest sample size (13 participants), the evaluation provided meaningful insight into how such a tutor might integrate into authentic educational settings. Students valued the opportunity to ask questions freely and receive immediate, supportive responses. Teachers appreciated the clarity and consistency of the tutor's output, especially for reinforcing or rephrasing core concepts. In both groups, the tutor was seen not as a replacement for teaching but as a complementary support — available on demand and reliably focused on the curriculum [15, 26]. This consistently constrained behavior was among the tutor's most appreciated features. Unlike generic chatbots, it neither drifted off-topic nor overwhelmed students with irrelevant or overly complex content. Instead, it followed a deliberately modest script: scaffold questions, explain concepts, encourage effort, and refrain from solving tasks outright. This made interactions feel safe, trustworthy, and pedagogically coherent.

5.2 From TAM to Trust and Pedagogy

While the Technology Acceptance Model (TAM) provided a solid analytical lens for measuring perceived usefulness and ease of use [7], feedback revealed important dimensions beyond its scope. One notable concern was *trust calibration*: some students were frustrated when the tutor refused to provide direct answers, while others relied on its output uncritically [18, 28]. This dual challenge is well-documented in human-AI interaction and must be addressed through transparency, uncertainty signaling, and strategies that scaffold reflection. Another theme that emerged was the tutor's influence on learning behavior. While many participants appreciated how it encouraged inquiry and reflection, some expressed concern that it might foster surface-level engagement or passive question-asking. These findings point to a need for integrating constructs such as *epistemic trust*, *learner agency*, and *pedagogical alignment* into future AI tutor evaluations.

5.3 Implications for SE and STEM Education

Although deployed in a general computing course, the tutor's design principles translate well to software engineering (SE) education, where learners work independently, face ill-structured problems, and require just-in-time conceptual and technical support [9, 23].

Key lessons include the value of *instructionally tuned prompting* [17], the importance of modular, tagged knowledge bases for precision and scalability, and the tutor's ability to scaffold learner autonomy through persistent, low-pressure guidance [19]. Its threaded memory and traversal logic enable adaptive support across tasks, while its refusal to provide direct solutions helps maintain the productive struggle essential to deep learning [4].

The tutor also demonstrates how AI can serve as a *metacognitive partner*, guiding learners through hypothesis testing, debugging, and design reasoning — all skills central to software engineering practice [1]. Importantly, logging student-AI interactions may support formative assessment by identifying misconceptions and guiding instructional interventions [14].

5.4 Reframing AI Tutoring: From Disruption to Enablement

Our findings support a shift from disruptive hype to grounded enablement. Echoing critiques by Reich and others [22], this tutor does not seek to transform education overnight. Instead, it extends access to individualized support in contexts where human help is limited. As such, it amplifies rather than replaces existing pedagogical practices.

Its novelty lies not in technical complexity but in deliberate constraint. It is safe, interpretable, curriculum-aligned, and easy to extend — qualities often overlooked in AI deployments. This makes it a viable template for educational use in schools, even those without deep technical infrastructure.

Rather than replacing classroom instruction, the AI tutor supports it — acting as a bridge between human mentorship and student independence. It thus signals a promising path to more inclusive, adaptive, and pedagogically grounded computing education.

6 Limitations

Although this exploratory study shows promising feasibility and acceptance of AI tutors in computing education, several limitations restrict the scope and generalizability of the findings.

First, the small sample (13 participants) and the mismatch between the test group (primarily male university students and pre-service teachers) and the intended target group (secondary school learners and teachers) limit external validity. Participants' higher digital literacy and maturity may have positively biased their experience, particularly regarding usability and interaction style [27].

Second, the AI tutor was evaluated as a prototype in a standalone web environment. Its integration into the RockStartIT platform has not yet been realized, which has affected contextual alignment and task continuity. Moreover, the evaluation sessions were short and one-off; without longitudinal data, we cannot assess how user trust, engagement, or learning strategies evolve over time [24].

Third, the system's current capabilities were constrained by platform limitations — such as lack of persistent memory, multimodal input, or fine-grained user modeling — hindering deeper personalization and task-specific support. In addition, no behavioral usage data (e.g., interaction logs, time-on-task) were collected, making it difficult to triangulate self-reports with actual learning behavior.

Finally, while most participants rated the tutor positively, the study did not explicitly assess trust calibration. As previous research shows, both over-reliance and under-reliance on AI pose educational risks [2]. Future evaluations must examine how learners engage critically with AI feedback and how trust can be scaffolded through interaction design.

7 Conclusion

This paper presented the design and pilot evaluation of a GPT-based AI tutor developed to support secondary-level learners in a digital software engineering course. By embedding the tutor into a real-world learning context and evaluating it with both students and teachers, we observed high levels of perceived usefulness, usability, and intention to use.

Our findings show that conversational AI, when carefully constrained and pedagogically structured, can enhance learner autonomy, engagement, and reflective thinking. The tutor's modular knowledge base — paired with instructionally aligned prompts — enabled scaffolded support while avoiding direct solution delivery, thus promoting inquiry rather than answer-seeking.

These results underscore AI tutors' potential to complement human instruction, particularly where real-time feedback is limited. The positive reception from students and educators suggests a viable role for such systems in future software engineering education.

Future work should focus on deployment with the intended target group in authentic classroom settings, where language needs and attention spans differ from adult learners. Longitudinal studies are needed to assess sustained engagement, trust calibration, and learning gains. Technical development will involve tighter integration with RockStartIT, synchronization with learner progress, and support for multimodal inputs (e.g., code snippets, diagrams).

Pedagogically, future versions should explore adaptive scaffolding techniques such as progressive hinting, reflective prompts, and feedback on misconceptions. Designing for trust — through strategies like uncertainty signaling or response justification — can further encourage critical thinking and epistemic agency. Improved onboarding and voice-enabled interaction may increase accessibility for younger learners or students with literacy challenges.

Ultimately, AI tutors like the one presented here offer a lightweight, scalable, and pedagogically grounded approach to augmenting software engineering education. Realizing this potential will require continued iteration, empirical validation, and close collaboration with educators — ensuring that AI becomes not a replacement, but a trusted partner in teaching and learning.

Acknowledgments

We thank the RockStartIT team and the participating students and teachers. This work was supported by funding from the pilot program Core Informatics at KIT (KiKIT) of the Helmholtz Association (HGF). This work was supported by the Topic Engineering Secure Systems of the Helmholtz Association (HGF) and supported by KASTEL Security Research Labs, Karlsruhe.

References

[1] Roger Azevedo and Vincent Aleven. 2013. *International handbook of metacognition and learning technologies*. Springer. 1000 pages. doi:10.1007/978-1-4419-5546-3

[2] Narges Bakhtiar and Sidney K. D'Mello. 2022. A Framework for Trust in Educational AI. *International Journal of Artificial Intelligence in Education* 32 (2022), 1–23.

[3] 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* (Milan, Italy) *(ITiCSE 2024)*. Association for Computing Machinery, New York, NY, USA, 394–400. doi:10.1145/3649217.3653543

[4] Michelene TH Chi, Marguerite Roy, and Robert G Hausmann. 2008. Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive science* 32, 2 (3 2008), 301–341.

[5] Larry Cuban. 2001. *Oversold and Underused: Computers in the Classroom*. Harvard University Press.

[6] Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*. 323–325. doi:10.1109/ICALT58122.2023.00100

[7] Fred D. Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13, 3 (9 1989), 319–340. doi:10.2307/249008 Publisher: Management Information Systems Research Center, University of Minnesota.

[8] Eduard Frankford, Clemens Sauerwein, Patrick Bassner, Stephan Krusche, and Ruth Breu. 2024. AI-Tutoring in Software Engineering Education, In Proceedings of the 46th International Conference on Software Engineering: Software Engineering Education and Training (Lisbon Portugal, 2024-04-14). *2024 IEEE/ACM 46th International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET)*, 309–319. doi:10.1145/3639474.3640061

[9] Philip Gross and Lucia Happe. 2022. Challenges and Opportunities of AI Tutors in Software Engineering Education. *Informatics in Education* (2022).

[10] Lucia Happe and Barbora Buhnova. 2022. Frustrations Steering Women Away From Software Engineering. *IEEE Software* 39, 4 (7 2022), 63–69. doi:10.1109/MS.2021.3099077

[11] Lucia Happe, Barbora Buhnova, Anne Koziolek, and Ingo Wagner. 2021. Effective measures to foster girls' interest in secondary computer science education. *Education and Information Technologies* 26, 3 (2021), 2811–2829. doi:10.1007/s10639-020-10379-x

[12] Lucia Happe and et al. 2025. Authentic Interdisciplinary Online Courses for Alternative Pathways into Computer Science. doi:10.1016/j.jss.2024.112240 Submitted to EAIT (2023).

[13] 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* (Chicago, IL, USA) *(ICER '23)*. Association for Computing Machinery, New York, NY, USA, 93–105. doi:10.1145/3568813.3600139

[14] Wayne Holmes et al. 2019. Artificial intelligence in education: Promises and implications for teaching and learning. *OECD Publishing* (2 2019). https://oro.open.ac.uk/60255/

[15] Kenneth Holstein and et al. 2019. Student perceptions of AI tutors in the classroom. In *CHI '19*.

[16] Greg Kestin, Kelly Miller, Anna Klales, Timothy Milbourne, and Gregorio Ponti. 2025. AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports* 15, 1 (June 2025). doi:10.1038/s41598-025-97652-6

[17] Daniel Lee and Edward Palmer. 2025. Prompt engineering in higher education: a systematic review to help inform curricula. *International Journal of Educational Technology in Higher Education* 22, 1 (2025), 7.

[18] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (3 2004), 50–80. https://www.engineering.uiowa.edu/~csl/publications/pdf/leesee04.pdf

[19] Raymond Lister et al. 2004. A multi-national study of reading and tracing skills in novice programmers. *SIGCSE Bulletin* 36, 4 (6 2004), 119–150. https://opus.lib.uts.edu.au/bitstream/10453/4126/3/2004000904.pdf

[20] Rose Luckin, Wayne Holmes, Marc Griffiths, and Laurie Forcier. 2016. Intelligence Unleashed: An Argument for AI in Education. https://oro.open.ac.uk/50104/

[21] Tanya Nazaretsky, Mutlu Cukurova, and Giora Alexandron. 2022. An instrument for measuring teachers' trust in AI-based educational technology. In *LAK22: 12th international learning analytics and knowledge conference*. 56–66.

[22] Justin Reich. 2020. Failure to Disrupt: Why Technology Alone Can't Transform Education. *Harvard University Press* (2020).

[23] Anthony Robins. 2010. Learning edge momentum: A new account of outcomes in CS1. *Computer Science Education* 20, 1 (3 2010), 37–71.

[24] Jeremy Roschelle, Roy Pea, Christopher Hoadley, Douglas Gordin, and Barbara Means. 2000. Changing How and What Children Learn in School With Computer-Based Technologies. *The Future of Children* 10, 2 (2000), 76–101. https://telearn.hal.science/hal-00190610/document

[25] Paul Saettler. 1990. *The Evolution of American Educational Technology*. Information Age Publishing.

[26] R. Winkler and M. Söllner. 2020. Unleashing the potential of chatbots in education: A TAM-based analysis. In *Proceedings of ICIS*.

[27] Yong Xu and Ming Li. 2022. AI in the Classroom: Challenges and Opportunities. *Educational Technology Research and Development* 70 (2022), 123–140.

[28] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos (Eds.). ACM, 1–12.

[29] Ying Zhu and Philip J. Guo. 2021. How AI Tutors Shape Student Reasoning in SE. In *ACM SIGCSE*.