



How Far Are We? The Triumphs and Trials of Generative AI in Learning Software Engineering

Rudrajit Choudhuri , Dylan Liu, Igor Steinmacher, Marco Gerosa, Anita Sarma



Introduction



- Generative AI (genAI) is revolutionizing SE
- However, uncertainty exists in how these tools can be leveraged in education
- Conversational agents has been shown to be useful for students
- GenAI has also been explored in this context:
 - focused on solving introductory CS problems
 - focused on programming

How can genAI tools be leveraged in supporting students in SE tasks ?

- that demand task-specific, contextualized assistance.

Research Questions (RQs)

RQ1

How effective is genAI in helping students in SE tasks?

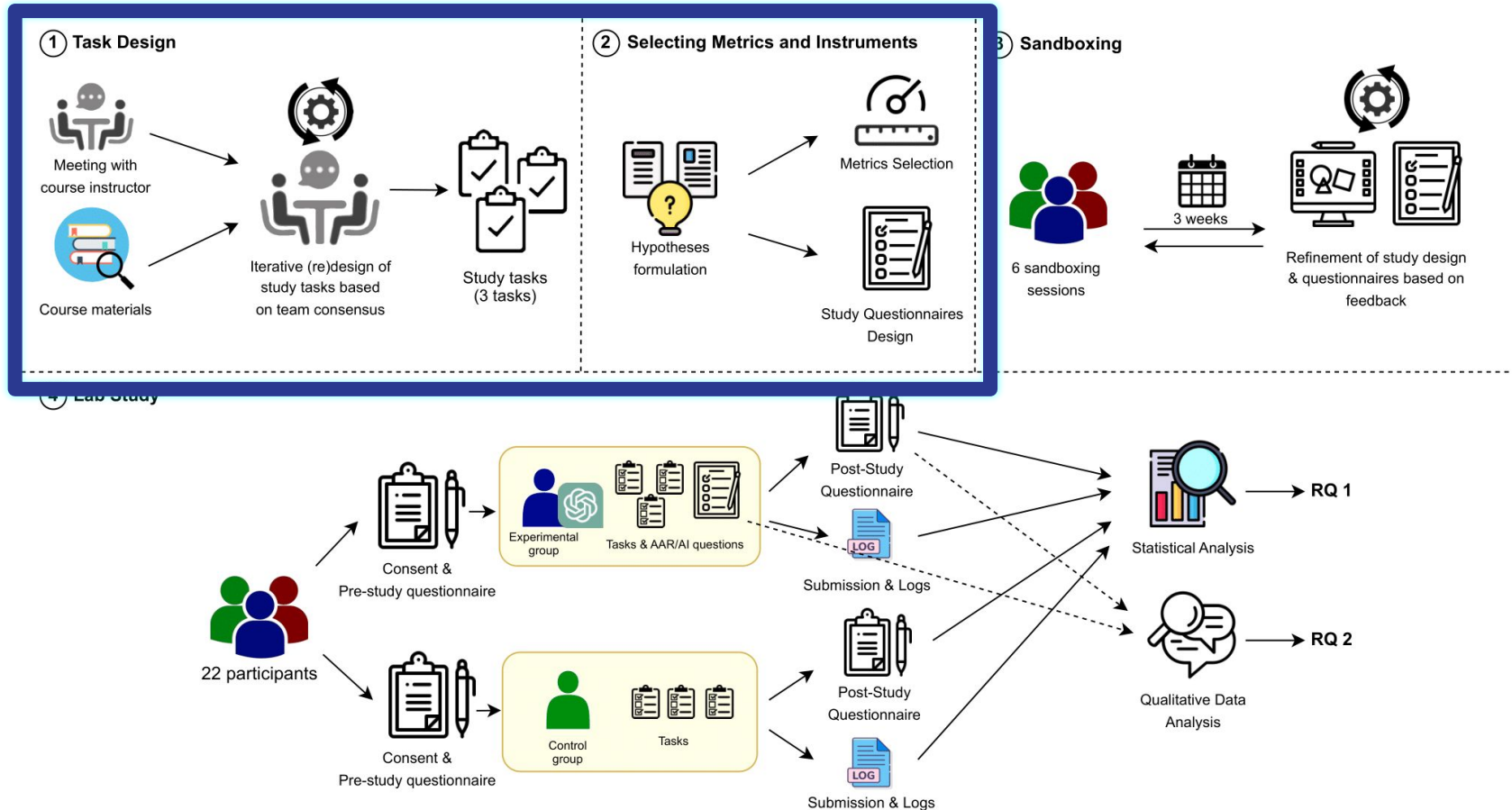
RQ2

What are the current pitfalls in genAI in helping students with SE tasks?

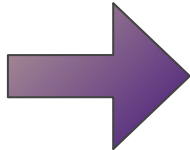
Between-subjects study (N=22) with students enrolled in SE courses at our university

- *Experimental*: ChatGPT (GPT-4) vs. *Control*: Non-GenAI resources

Method



Task Design



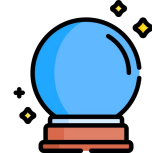
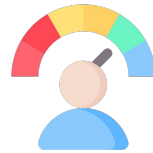
Fixing code functionalities involving third party APIs

Removing code smells



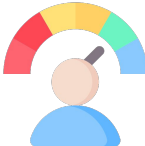

Contributing changes to a GitHub repository via pull requests

Iteratively designed and reviewed based on
instructor's input

RQ1: Effectiveness in helping students with SE tasks?



Selecting Metrics and Instruments (RQ1)

	Construct	Metrics & Instruments	
	Cognitive Load	NASA TLX [1]	<p>H1: <i>Participants using ChatGPT for the tasks perceive lower cognitive load than those using alternate resources.</i></p>
	Productivity	Task Correctness & Time to Complete [2]	<p>H2: <i>ChatGPT positively impacts participants' productivity.</i></p>
	Self-efficacy	Self-efficacy questions [3]	<p>H3: <i>ChatGPT promotes participants' self efficacy.</i></p>
	Continuance Intention	Direct likelihood questions [4]	<p>Part of the post study questionnaire</p>

Results (RQ1): Effectiveness (Cognitive Load)

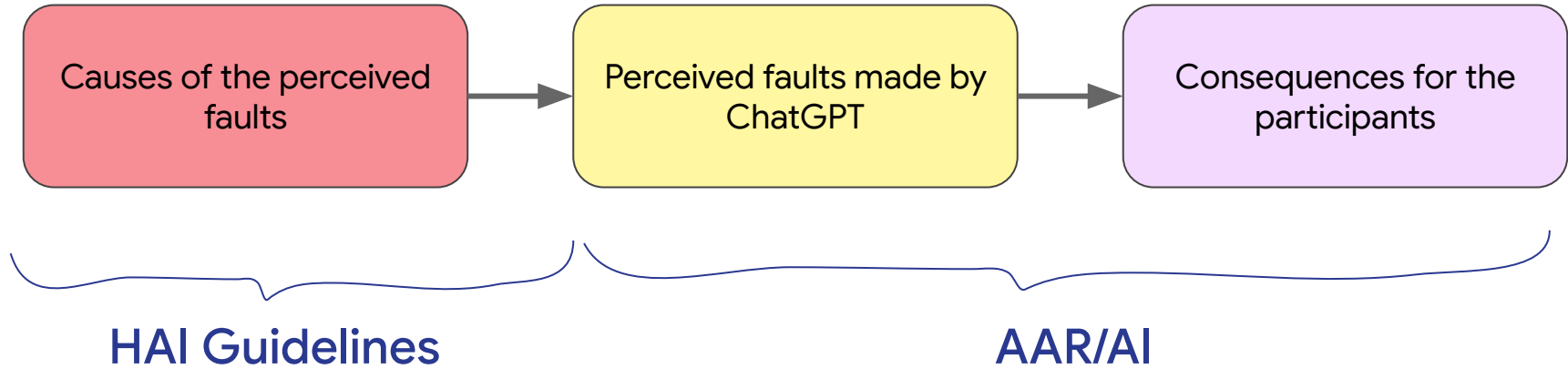
	NASA TLX					Frustration
	Mental	Physical	Temporal	Performance	Effort	
Estimate	47	51.5	64.5	45.5	45.5	101
p-value	0.388	0.557	0.817	0.339	0.337	0.008***
Cliff's delta(δ)	-0.223	-0.149	0.066	-0.248	-0.248	0.669
Median values for each group						
Experimental	15	1	15	9	14	14
Control	14	3	15	12	14	9

Higher frustration levels among participants using ChatGPT.

"...it misinterpreted my questions, was REALLY slow, and didn't account for errors. It was hopeless (PT-7)"

H1 is not supported: Participants using ChatGPT did not perceive statistically significant lower cognitive load than those using alternate resources.

RQ2: Pitfalls in helping students with SE tasks?



Selecting Metrics and Instruments (RQ2)

After Action Review for AI (AAR/AI)

(pronounced “arf-eye”, short for AAR for AI)

- Standardized AI assessment process to help end users find AI faults [5]
- Recent member of the After-Action Review [6] family,
 - Devised by the U.S. military in the 1970s as a facilitated debriefing method
 - Used for decades and has been successfully adapted to different domains [7, 8]
- Integrating AAR/AI helps end users uncover significant number of faults with greater precision [5, 9]

Results (RQ2): Pitfalls (perceived AI faults)

F1: Limited advice on niche topics

“for anything that wasn't super standard, ChatGPT struggled to easily give useful answers (PT-1)”

F2: Inability to comprehend the problem

“it identified non-problems as problems and missed actual ones and didn't do the thing I wanted it to do despite giving it context (PT-6)”

F3: Incomplete assistance

“...[ChatGPT] did not give me answers on how to solve the whole task (PT-11)”.

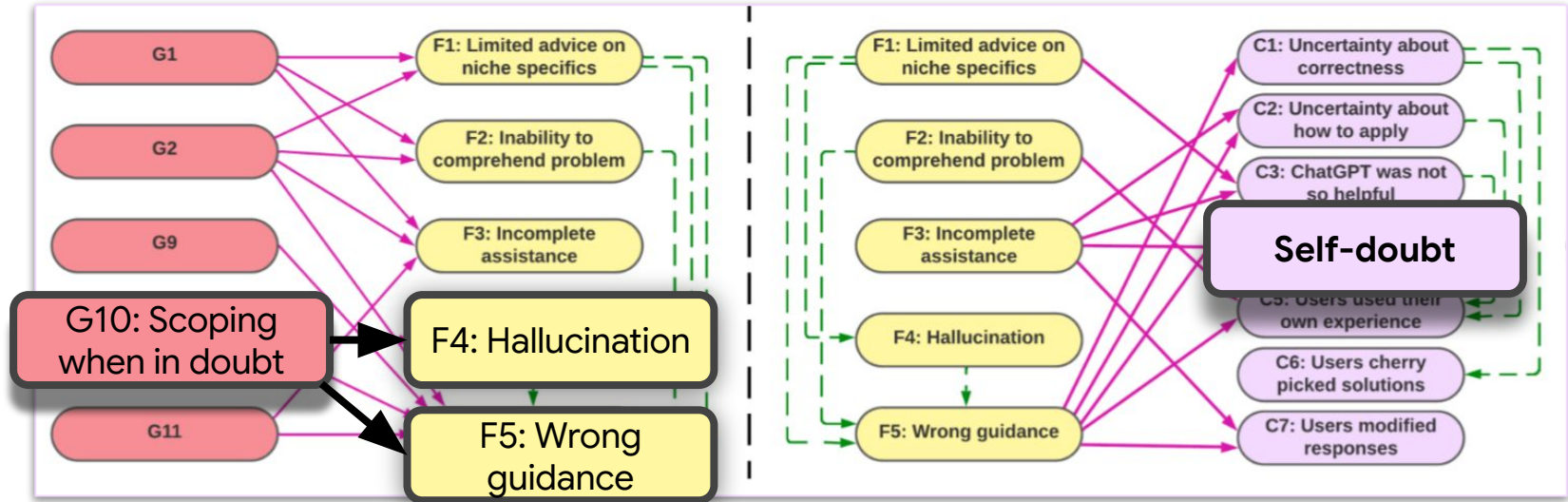
F4: Hallucination

“made up parameters for functions that were unfamiliar (PT-4)”

F5: Wrong guidance

“It couldn't figure out test case 3 and kept telling me to check my drivers...without realizing there were missing imports (PT-8)”.

Results (RQ2): Pitfalls (perceived AI faults, causes, consequences)



Participants reported that ChatGPT violated 5 of the 18 HAI guidelines.

These faults had consequences on the participants.

Cascading faults: “ChatGPT did not have as much knowledge . . . and confidently told me incorrect ways to ‘fix’ my code (PT-9)”

Self-doubt: “I got confused over its suggestions, ... likely I may have asked something wrong (PT-2)”

Conclusion

Task 1: Fixing Code Functionalities

Task 2: Removing Code Smells

Task 3: Contributing changes via PRs



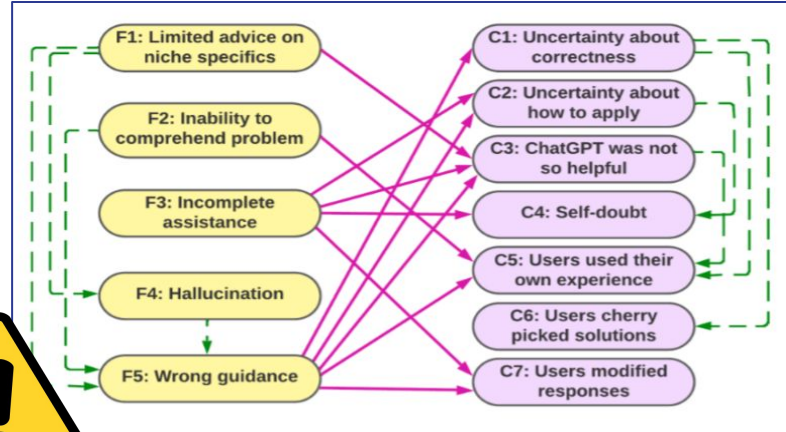
Cognitive Load ↓ ✗



Productivity ↑ ✗



Self-Efficacy ↑ ✗



“For anything that wasn’t super standard, ChatGPT struggled to easily give useful answers (PT-1)”

- Expert developers can navigate this, but novices might struggle or learn incorrect practices.
- Necessary to :
 - customize genAI with pedagogical scaffolds to support students
 - follow iterative participatory approach in future genAI design



Check out our
paper!

Thank You!
Questions?

choudhru@oregonstate.edu

NAU NORTHERN ARIZONA
UNIVERSITY



Oregon State
University

References

- [1] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183
- [2] Frank F Xu, Bogdan Vasilescu, and Graham Neubig. 2022. In-IDE code generation from natural language: Promise and challenges. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 31, 2 (2022), 1–47.
- [3] Igor Steinmacher, Tayana Uchoa Conte, Christoph Treude, and Marco Aurélio Gerosa. 2016. Overcoming open source project entry barriers with a portal for newcomers. In *Proceedings of the 38th International Conference on Software Engineering*. 273–284.
- [4] MY Park and KH Chung. 2011. The antecedents and consequences of user satisfaction in virtual community: Focused on college students. *Korean Research Academy of Distribution and Management Review* 14, 1 (2011), 77–99.
- [5] Jonathan Dodge, Roli Khanna, Jed Irvine, Kin-Ho Lam, Theresa Mai, Zhengxian Lin, Nicholas Kiddle, Evan Newman, Andrew Anderson, Sai Raja, et al. 2021. After-action review for AI (AAR/AI). *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11, 3-4 (2021), 1–35.
- [6] John E Morrison and Larry L Meliza. 1999. Foundations of the after action review process. Technical Report. Institute for Defense Analyses Alexandria Va.
- [7] Andrew W Ishak and Elizabeth A Williams. 2017. Slides in the tray: How fire crews enable members to borrow experiences. *Small Group Research* 48, 3 (2017), 336–364.
- [8] Taylor Lee Sawyer and Shad Deering. 2013. Adaptation of the US Army’s after action review for simulation debriefing in healthcare. *Simulation in Healthcare* 8, 6 (2013), 388–397
- [9] Roli Khanna, Jonathan Dodge, Andrew Anderson, Rupika Dikkala, Jed Irvine, Zeyad Shureih, Kin-ho Lam, Caleb R Matthews, Zhengxian Lin, Minsuk Kahng, et al. 2022. Finding AI’s faults with AAR/AI: An empirical study. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 12, 1 (2022), 1–33.

Backup Slides

Table 2: AAR/AI steps and our adaptations. The Empirical context column explains how we realized the method in our study. Steps 3 to 6 were “inner loop” questions we repeated for all three tasks.

AAR/AI Steps	AAR/AI in our Empirical context
1. Defining the rules: How are we going to do this evaluation? What are the details regarding the situation?	We briefed the participants about the study details and how we were going to do the evaluation. Then we stated: “You will be given a questionnaire before and after each task. Please be detailed in your responses as that will help us evaluate ChatGPT’s performance.”
2. Explaining the objectives of the AI agent: What is the AI’s objective for this situation?	We oriented the participants about the primary objective of ChatGPT by stating, “The primary objective of ChatGPT will be to assist you by providing contextual, disambiguous, and correct information.”
Inner Loop	
3. Reviewing what was supposed to happen: What did the evaluator intend to happen?	We asked “What do you think should happen when you use ChatGPT for this task?” The participants chose between: It will (provide (all/some))/(not provide any) useful information I need to complete the task.
4. Identify what happened: What actually happened?	The participants did a task, then we asked “What actually happened when you used ChatGPT for this task?” The participants chose between: It (provided (all/some))/(did not provide any) useful information I need to complete the task.
5. Examine why it happened: Why did things happen the way they did?	We asked “Why do you think ChatGPT behaved this way?”
6. Formalize learning (end inner loop): What changes would you make in the decisions made by the AI to improve it?	We asked two questions: “To what extent did you modify ChatGPT’s responses for solving the task?” The participants chose between: Did not modify at all/Modified (slightly/significantly). Then, we asked them to “Briefly explain why?”
End Inner Loop	
7. Formalize learning: What went well, what did not go well, what could be done differently next time?	We asked three questions: “What went well?”, “What did not go well?”, “What could be done differently next time?”

1
INITIALLY

Make clear what the system can do

Help the users understand what the AI system is capable of doing.

2
INITIALLY

Make clear how well the system can do what it can do.

Help the user understand how often the AI system may make mistakes.

Microsoft's 18 HAI guidelines recommend how AI systems should behave upon initial interaction, during regular interaction, when they're inevitably wrong, and over time.

Here are some of them:

7
WHEN WRONG

Support efficient invocation.

Make it easy to invoke or request the AI system's services when needed.

8
WHEN WRONG

Support efficient dismissal.

Make it easy to dismiss or ignore undesired system services.

9
WHEN WRONG

Support efficient correction.

Make it easy to edit, refine, or recover when the AI system is wrong.

10
WHEN WRONG

Scope services when in doubt.

Engage in disambiguation or gracefully degrade the AI system's services when uncertain about a user's goals.

11
WHEN WRONG

Make clear why the system did what it did.

Enable the user to access an explanation of why the AI system behaved as it did.

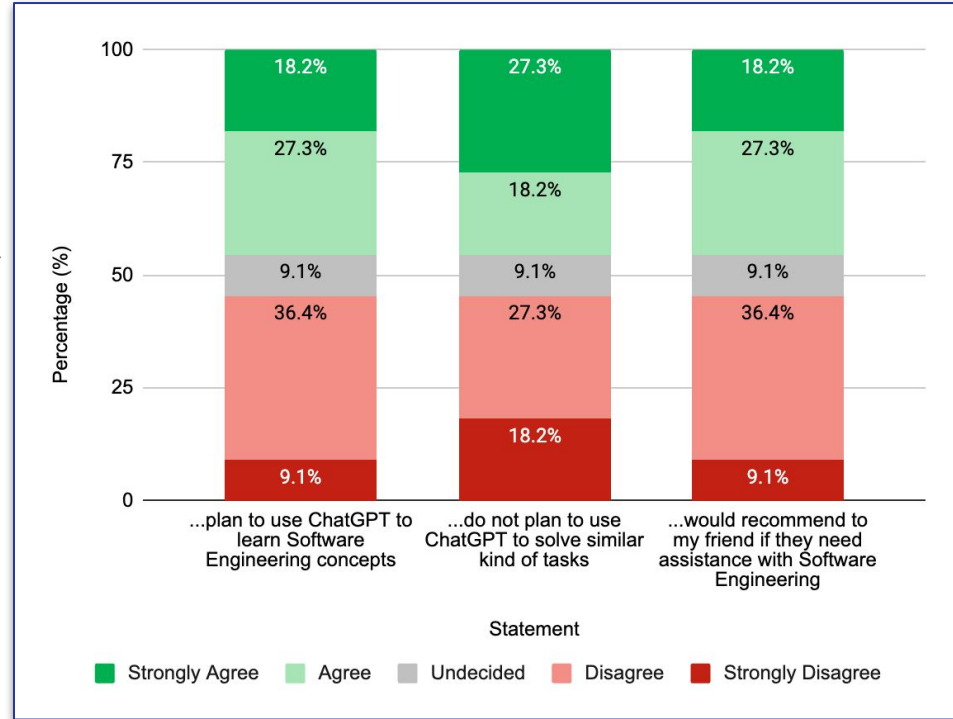
Results (RQ1): Effectiveness

Polarized continuance intention: while one half of the participants intended to use genAI for SE, the others equally resisted:

"I would have liked to be able to ask someone knowledgeable in Python about [task 1] (PT-11)"



"I could not rely on [ChatGPT] to tell me when functions exist or not (PT-1)"



Results (RQ2): Pitfalls (AI faults, their causes & consequences)

F1: Limited advice on niche topics.

ChatGPT struggled to provide advice on topics specific to a niche (e.g., a domain, a library, or a concept): *“for anything that wasn't super standard, ChatGPT struggled to easily give useful answers (PT-1)”*

F2: Inability to comprehend the problem.

ChatGPT couldn't always comprehend participants' goals or problems. *“It identified non-problems as problems and missed actual problems”* and didn't *“do the thing you want it to do despite giving it context (PT-6)”*

F3: Incomplete assistance.

ChatGPT often provided incomplete/partially correct assistance even when it was able to grasp the problem *“...it did not give me answers on how to solve the whole task (PT-11)”*.

F4: Hallucination.

ChatGPT hallucinated, creating false answers when it didn't know the correct solution and *“made up parameters for functions that were unfamiliar” (PT-4)*.

F5: Wrong guidance.

In addition to hallucinating, there were other instances where ChatGPT gave wrong guidance, or *“incorrect ways to fix [problems] (PT-9)”*. For example, when it could not comprehend the problem (F2), PT-8 was facing, it gave a piece of incorrect advice: *“It couldn't figure out test case 3 and kept telling me to check my drivers...without realizing there were missing imports (PT-8)”*.