

# “I’m Not Reading All of That”: Understanding Software Engineers’ Level of Cognitive Engagement with Agentic Coding Assistants

Carlos Rafael Catalan  
Samsung R&D Institute Philippines  
Manila, Philippines  
c.catalan@samsung.com

Patricia Nicole Monderin  
Samsung R&D Institute Philippines  
Manila, Philippines  
p.monderin@samsung.com

Lheane Marie Dizon  
Samsung R&D Institute Philippines  
Manila, Philippines  
lm.dizon@samsung.com

Emily Kuang  
York University  
Toronto, Canada  
ekuang@yorku.ca

## Abstract

Over-reliance on AI systems can undermine users’ critical thinking and promote complacency, a risk intensified by the emergence of agentic AI systems that operate with minimal human involvement. In software engineering, agentic coding assistants (ACAs) are rapidly becoming embedded in everyday development workflows. Since software engineers (SEs) create systems deployed across diverse and high-stakes real-world contexts, these assistants must function not merely as autonomous task performers but as Tools for Thought that actively support human reasoning and sensemaking. We conducted a formative study examining SEs’ cognitive engagement and sensemaking processes when working with an ACA. Our findings reveal that cognitive engagement consistently declines as tasks progress, and that current ACA designs provide limited affordances for reflection, verification, and meaning-making. Based on these findings, we identify concrete design opportunities leveraging richer interaction modalities and cognitive-forcing mechanisms to sustain engagement and promote deeper thinking in AI-assisted programming.

## ACM Reference Format:

Carlos Rafael Catalan, Lheane Marie Dizon, Patricia Nicole Monderin, and Emily Kuang. 2026. “I’m Not Reading All of That”: Understanding Software Engineers’ Level of Cognitive Engagement with Agentic Coding Assistants. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (CHI-Tools for Thought Workshop)*. ACM, New York, NY, USA, 7 pages. <https://doi.org/XXXXXXX.XXXXXXX>

## 1 Introduction

Recent advances in generative AI systems have brought about immense productivity gains in human-AI co-creative work. In software engineering in particular, agentic coding assistants (ACAs) such as Cline [11], Claude Code [4], and Codex [34] increasingly act as autonomous collaborators that generate, modify, and reason

about code [19, 45]. While these systems demonstrably accelerate development workflows, their growing agency raises critical questions about how they shape human cognition during complex problem-solving tasks.

Prior research has shown that over-reliance on AI systems can negatively affect core cognitive capabilities, such as decision-making [2, 37], critical thinking [12], and problem-solving [12]. These capabilities are especially crucial in software engineering, where developers must evaluate trade-offs, reason about correctness, and anticipate failure modes. At the same time, large language models (LLMs) that power these ACAs remain prone to hallucinations [16] and algorithmic biases [30].

Framed through the lens of AI as “Tools for thought,” these tensions point to a design challenge rather than a purely technical one: how might ACAs be designed and used in ways that protect and augment human cognition, rather than displacing it? Understanding this requires moving beyond performance metrics to examine how developers cognitively engage with ACAs, how they make sense of AI-generated suggestions, and how they reflect on and evaluate these interactions during real work.

In this workshop contribution, we investigate software engineers’ cognitive engagement when working with agentic coding assistants to inform design and usage strategies for GenAI as tools for thought. Specifically, we ask: **RQ1**: *How cognitively engaged with the task are software engineers when working with agentic coding assistants?*, and **RQ2**: *How did they recall, understand, analyze, and evaluate different aspects of their interaction with an ACA?*

We conducted a formative study with four software engineers whose professional experience ranged from less than 1 year to more than 10 years. Participants performed a single code generation task with an ACA, specifically Cline [11]. These tasks were designed and categorized using Bloom’s Taxonomy as a theoretical framework for examining cognitive engagement [8] Following the code generation task, participants self-reported their level of cognitive engagement through a survey assessing their reasoning, attention, and reflection during the task. Our findings reveal two key patterns. First, participants’ cognitive engagement declined as tasks progressed. Second, they were primarily focused on achieving correct outputs, paying little attention to the underlying process that produced them. Together, these behaviors led to shallow engagement, limiting participants’ ability to accurately recall, understand, analyze, and evaluate important details of the task. To address this

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*CHI-Tools for Thought Workshop, Barcelona, Spain*

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2026/04

<https://doi.org/XXXXXXX.XXXXXXX>

decline in engagement, we propose design considerations for ACAs aimed at sustaining cognitive engagement and encouraging critical thinking beyond merely reaching correct solutions.

## 2 Background

### 2.1 Bloom’s Taxonomy as Measurement of Cognitive Engagement

Bloom’s Taxonomy conceptualizes cognition as a hierarchy of cognitive processes that vary in complexity and depth, progressing from basic forms of information processing to more advanced forms of reasoning and knowledge construction [3, 8]. At the lower end of this hierarchy, processes such as remembering and understanding involve the recognition and interpretation of information, whereas higher-level processes, such as analyzing and evaluating, require individuals to synthesize that information to reason about structures and relationships [21]. In this taxonomy, higher-order cognition depends on the successful operation of foundational processes such that complex reasoning and judgment presuppose accurate recall and meaningful comprehension of relevant information, as well as coherent mental representations of task elements and their relationships [15].

This framework provides a conceptual basis for examining cognitive engagement in interactions with ACAs by characterizing engagement in terms of distinct cognitive processes. As ACAs increasingly assume planning and decision-making functions, task success alone offers limited insight into the underlying cognitive engagement of users. As such, cognitive engagement must instead be examined through the cognitive operations users employ while interacting with an ACA. Within this framework, levels such as remembering, understanding, analyzing, and evaluating correspond to qualitatively different forms of engagement, ranging from recall of interaction information to higher-order reasoning about the behavior and reliability of ACAs. This perspective allows us to assess both how cognitively engaged users are when interacting with ACAs and how that engagement is distributed across recall, comprehension, analytical reasoning, and evaluative judgment.

### 2.2 Self-Report Surveys for Cognitive Engagement

Self-report surveys serve as a reliable measurement tool for cognitive engagement [18], because it uses participants’ verbal responses to assess their current conscious cognitive state [36]. These surveys have been used extensively in the education domain to measure students’ motivations and learning goals [13, 14]. However, self-reporting is prone to memory biases. This is particularly true when the survey is administered at a later point in time and requires individuals’ recall of key events from their autobiographical memory [36]. For our use case, we use Bloom’s Taxonomy as the framework for developing the questions for our survey. Specifically, we ask questions that would understand how well software engineers recalled, understood, analyzed, and evaluated various aspects of their interaction with Cline. To minimize memory biases, we administer this survey immediately after we ask them to perform a code generation task.

### 2.3 Cognitive Load Theory

Cognitive Load Theory suggests three types of cognitive demands on users during learning activities: intrinsic, extraneous, and germane load [5]. Intrinsic load pertains to the amount of existing cognitive resources the user needs to use to understand complex information. Extraneous load refers to the cognitive resources the user needs to expend to comprehend irrelevant information resulting from poor material design. Germane load refers to the allocation of cognitive resources towards problem solving and metacognition [22, 32, 40]. To enhance users’ deep understanding of the learning material, it must be designed in a way that minimizes extraneous load, optimizes intrinsic load, and promotes germane load [10, 28, 29].

## 3 Method

We conducted a user study to understand SEs’ level of engagement with an ACA’s output. We recruited four participants from a large software company in the Philippines. Each participant (P) represents a different category of professional SE experience: P1: less than 1 year, P2: 1-5 years, P3: 6-10 years, P4: more than 10 years. For the ACA, we selected Cline due to its agentic abilities of comprehensive planning and discussion with the SE, as well as invocation of external tools when executing the task [6]. Cline is also the tool that participants were most familiar with. The agent is also open source [11], affording us the flexibility to implement prototypes that encourage active user engagement for future work.

### 3.1 Code Generation Task

After going through the consent process, the moderator explained the study tasks. Each participant was presented with a laptop where the Visual Studio Code integrated development environment (IDE) is open, and the Cline agent is pre-loaded with the prompt for a Code Generation task:

Can you write a <programming language of your choice> script that checks all the Excel files in the folder and finds the one with a “dashboard” sheet? In the dashboard sheet, copy the values from column C to E. Then, generate another workbook that copies all the data from the current workbook, and names the new workbook’s sheet to whatever the name of the current workbook is.

This prompt was retrieved from DevGPT, a curated dataset of conversations between SEs and ChatGPT [46].

Before beginning the code generation task, the moderator explained to the participants that they are afforded as much time as needed in interpreting the prompt and what they expect to be the correct output. They were also briefed on Cline’s possible clarifying questions during the interaction, and that they should carefully analyze and use their best judgment when responding. Lastly, they were allowed to go back to any part of their conversation with Cline during the duration of the code generation task.

Once they understood the instructions, participants began the task by first indicating the programming language they were most comfortable working with before prompting Cline. Cline first began in ‘Plan’ mode to provide them with an approach to solving the

task. Once they thought the plan was sufficient, they then triggered 'Act' mode to generate the code. During the 'Plan' or 'Act' modes, Cline may ask them some clarifying questions. The task ended once "Start New Task" appeared; they may do some reviewing before notifying the moderator that they were done. During the entire task, the moderator was present in the room to observe and record the participants' behavior. They were also allowed to think out loud, even though the moderator did not explicitly state the protocol.

### 3.2 Survey Task

Before proceeding with the survey, participants were instructed that they were not allowed to go back to the IDE. The survey was divided into five parts: (1) Years of Experience, (2) Recall, (3) Understanding, (4) Analyzing, and (5) Evaluation. The survey was deployed on Qualtrics. Parts 2 to 5 contained questions aimed at measuring how well they **recalled, analyzed, understood, and evaluated** certain aspects of their interaction with Cline, following Bloom's Taxonomy.

### 3.3 Data Analysis

We first cross-referenced each participant's responses to the recall questions with their corresponding working directory during the code generation task to determine if they correctly recalled various aspects of their interaction (e.g., the name of the folder they were working on). Then, two authors conducted a thematic analysis on the participants' short-answer responses and the moderator's observational notes [33]. From this analysis, we identified two key patterns, which we discuss in the next section.

## 4 Findings

We divide the interaction between the SE and Cline into three phases: planning, execution, and evaluation. The **planning** phase involves the SE understanding the prompt and the expected output, Cline creating a plan for the task, and the SE reviewing the plan. The **execution** phase involves Cline generating the actual code, running it, and evaluating it. The **evaluation** phase involves the SE evaluating both the generated source code and Cline's evaluation of that source code.

### 4.1 SEs' Cognitive Engagement Declines as the Task Progresses

In the context of our study, we observed a clear downward trend in the SEs' level of cognitive engagement as the task progressed, which provides some formative insights for RQ1. Participants devoted the greatest cognitive effort during the planning phase, focusing on guiding the agent toward producing correct outputs. However, engagement dropped during the execution phase, where the volume of information presented by Cline often led to disengagement. As a result, during the evaluation phase, SEs allocated minimal cognitive resources to verifying the output and largely neglected reviewing the underlying process that generated it.

*4.1.1 Software Engineers Allocated Most of Their Cognitive Resources Towards the Planning Phase.* We found that all participants allocated substantial cognitive resources in the Planning Phase,

as evidenced by our observations that they were all actively engaged with the task at the beginning, especially when trying to comprehend the prompt. For example, P3 asked out loud: *What is the context of this?* P1, P2, and P4 took the time to verify the files in the working directory before prompting Cline.

Once Cline generated the plan, most participants went back through the conversation to read it. P1, P2, and P4 actively read the conversation by going back and forth between it and the generated plan, or slowly scrolling through it. This is in contrast to P3, who just skimmed the conversation to save time. We hypothesize that SEs attributed importance to tasks in relation to planning and orchestration of the agent to ensure that the agent understood the requirements to increase the probability of correctly performing the task. This may be a form of germane cognitive load, which led the participants to be meticulous in answering Cline's clarifying questions and reading through the conversation.

*4.1.2 Information Overload During Cline's Execution Phase.* During Cline's execution phase, we observed that participants' engagement decreased while Cline was executing its process of code generation and tool invocation. Some of them (P2, P3) looked away from the laptop and towards the other parts of the room, and only when Cline provided results or clarification did the participants go back to the task. However, some participants did not necessarily engage with the results, as evidenced by P4's comment: *"I'm not reading all of that"*. We also observed that other participants (P2, P3) were quick to prompt Cline for the next steps, even though there was still a considerable amount of information on the screen that had just been generated. Participants perceived this "information dump" as extraneous cognitive load, which overwhelmed their attention and reduced meaningful engagement with the task. Moreover, Cline's reliance on text-only communication may have further amplified this extraneous load by requiring users to parse dense, unstructured information without additional visual or interactive support.

*4.1.3 Participants were More Concerned with Output Evaluation Rather than Process Evaluation.* In the evaluation phase, our findings show that the four participants engaged more with the output than the process, as evidenced by the following responses: (P1): *"It generated my desired output"*; (P2): *"I trust Cline"*, (P3): *"It worked"*, and (P4): *"I think what cline wrote was well-documented and readable enough, so I felt that changes were not necessary"*. Further evidence for this finding was also observed with P4. During the task, he reflected on his expectations and evaluated that there were inconsistencies in the output, so he prompted Cline with clearer instructions until it gave him the correct results.

Once Cline successfully generated the correct Excel file, participants chose not to make further changes to the code. In this task, the Excel file served as the final output, while the code constituted the underlying process. Both required intrinsic cognitive effort to evaluate; however, verifying the output demanded substantially fewer cognitive resources than reviewing the code itself. When the output appeared correct, participants treated the task as complete (as observed in P2 and P3). Only when errors occurred did they engage in deeper process-level evaluation and revision (as observed in P4). These behaviors suggest that software engineers adopt a *greedy allocation strategy* for cognitive resources, prioritizing the

lowest-effort checks that confirm task completion, and engaging in deeper reasoning only when surface-level validation fails.

## 4.2 Participants Only Recalled, Understood, Analyzed, and Evaluated the "Happy Path"

We offer some preliminary insights for RQ2. In the context of our study, participants appeared to recall, understand, analyze, and evaluate primarily the "happy path," that is, the sequence of steps leading to the correct output. This tendency to focus on the output rather than the process may have led participants to overlook potentially critical issues, which was reflected in the survey results. In terms of recall, **none** of the participants were able to correctly answer how many functions the generated script had. For understanding, only **half** of them understood what the first function did and could reliably provide a concise summary for it. Likewise, only **half** were able to adequately analyze the code and felt confident that it could handle edge cases.

These responses suggest that participants' evaluations may have been too narrow in scope, potentially overlooking unexpected or edge-case scenarios. Consider, for instance, a scenario in which an ACA is powered by an LLM susceptible to vulnerabilities such as poisoning and backdoor attacks [1, 39]. In such cases, SEs may need to recall, understand, analyze, and evaluate even seemingly minor aspects of their task to reduce the risk that malicious code goes undetected.

## 5 Design Opportunities for ACAs

### 5.1 Sustain Engagement by Communicating Beyond Text

During the ACA's execution phase, participant engagement appeared to be at its lowest. We hypothesize that the volume of text generated in real-time by the ACA may exceed what SEs can reasonably process. This limitation may be partly attributable to current ACAs relying solely on text as their communication medium. Faced with large amounts of streaming text, SEs may be inclined to skim rather than read carefully, potentially causing them to miss critical information relevant to their task.

To mitigate this, we propose to add visualization and alternative modalities, such as voice, to ACAs. Visualizations have long been an indispensable tool for human-centered design due to their ability to synthesize and communicate complex information [42] for the users [31, 43]. We envision an ACA that communicates its plans and reasoning in a presentable format, such as flowcharts, graphs, and mind maps. Previous work has also shown state-of-the-art voice AI systems that capture natural, human-like speech synthesis act as favorable social cues that elicit positive social responses from users [7, 26, 27, 38]. Studies by Liew et al. [25] and Wang et al. [41] showed that the implementation of multiple AI voice technologies in learning systems resulted in significantly lower perceived cognitive load of the information presented, and improved retention and recall of the information by the user.

### 5.2 Cognitive Forcing Designs to Encourage Critical Thinking

In our study, participants' responses centered on the happy path, with few instances of extended recall, understanding, analysis, and evaluation of the ACA's process and output. This suggests some over-reliance on the ACA for handling all possible edge cases. This observation aligns with the dual-process cognitive theory of *System 1* thinking, where humans employ shortcuts (in this case, analyzing the happy path) in decision making to minimize use of cognitive resources [20, 44].

Participants' prior experience with Cline and understanding of Cline's capabilities may have encouraged them to stay on *System 1* thinking. Designing systems that encourage users to switch to *System 2* or analytical thinking is a long-standing challenge in human-AI collaboration research because it is more effort-intensive on the part of the user. Previous works on *cognitive forcing designs* by Bućinca et al. [9] and Ghosh et al. [17] show these to be a promising approach to encourage *system 2* thinking.

Thus, we argue for such designs to be extended to ACAs. These designs are interventions that are applied during the AI's decision-making to disrupt its reasoning. This, in turn, *forces* the user to perform analytical thinking in the task [24]. This "slowing down" of the AI's decision has been shown to greatly increase user accuracy of the assessment towards it [35]. This finding aligns with another work done by Kuang et al. [23] in the domain of usability testing. In their work, AI's suggestions for usability improvements were only shown after the user was able to critically analyze the usability test video demonstration. This resulted in an improved perception of efficiency and trust towards the AI suggestions as it aligned with their desire to initially draw from their own expertise to address the task, and simply use the AI's analysis as a verification tool [23].

## 6 Limitations, Future Work, and Conclusion

Our work contains several limitations. We plan to continue this study by recruiting more participants, and add a software engineering task that is much more open-ended that would reward cognitive engagement. We also want to have an even deeper understanding of cognitive engagement to strengthen our design considerations. In addition to self-report surveys, we also plan to involve eye-tracking software to determine points of attention or lack thereof.

In conclusion, we present a formative study examining SEs' level of cognitive engagement when interacting with ACAs for code generation tasks. We found a concerning decline in cognitive engagement with the task as the interaction with the ACA progressed, which led our participants to overlook critical details. To support the development of correct and safe real-world software, we propose design considerations for ACAs that sustain cognitive engagement and encourage critical thinking throughout the workflow, including multimodal interactions (e.g., visualizations and voice), and cognitive forcing designs. These design considerations are grounded in existing literature as well as our framing of ACAs as pair programmers. Drawing from established pair-programming practices, where collaboration is inherently multimodal and includes instructional scaffolding, we argue that ACAs should similarly support rich communication and intentionally provoke reflection rather than immediately providing solutions.

## References

- [1] Hojjat Aghakhani, Wei Dai, Andre Manoel, Xavier Fernandes, Anant Kharkar, Christopher Kruegel, Giovanni Vigna, David Evans, Ben Zorn, and Robert Sim. 2024. TrojanPuzzle: Covertly Poisoning Code-Suggestion Models. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, Los Alamitos, CA, USA, 1122–1140. doi:10.1109/SP54263.2024.00140
- [2] Sayed Fayaz Ahmad, Heesup Han, Muhammad Mansoor Alam, Mohd Rehmah, Muhammad Irshad, Marcelo Arraño-Muñoz, Antonio Ariza-Montes, et al. 2023. Impact of artificial intelligence on human loss in decision making, laziness and safety in education. *Humanities and Social Sciences Communications* 10, 1 (2023), 1–14.
- [3] Lorin W. Anderson and David R. Krathwohl (Eds.). 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, New York.
- [4] Anthropic. 2025. Claude Code. <https://claude.com/product/claude-code>. Accessed: Dec 2025.
- [5] Maria Bannert. 2002. Managing cognitive load—recent trends in cognitive load theory. *Learning and Instruction* 12, 1 (2002), 139–146. doi:10.1016/S0959-4752(01)00021-4
- [6] Nick Baumann. 2025. Plan smarter, code faster: Cline's plan act is the paradigm for Agentic Coding - Cline blog. <https://cline.bot/blog/plan-smarter-code-faster-clines-plan-act-is-the-paradigm-for-agentic-coding>
- [7] Maik Beege and Sascha Schneider. 2023. Emotional design of pedagogical agents: the influence of enthusiasm and model-observer similarity. *Educational technology research and development* 71, 3 (2023), 859–880.
- [8] Benjamin S. Bloom et al. 1956. *Taxonomy of Educational Objectives: The Classification of Educational Goals*. Longmans, Green.
- [9] Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (April 2021), 21 pages. doi:10.1145/3449287
- [10] Ruth C Clark and Richard E Mayer. 2023. *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*. John Wiley & sons.
- [11] Cline. 2025. Cline - AI Coding, Open Source and Uncompromized. <https://cline.bot/>. Accessed: Dec 2025.
- [12] Arvin V Duhaylungsod and Jason V Chavez. 2023. ChatGPT and other AI users: Innovative and creative utilitarian value and mindset shift. *Journal of Namibian Studies: History Politics Culture* 33 (2023), 4367–4378.
- [13] Noel Entwistle and Paul Ramsden. 2015. *Understanding student learning (routledge revivals)*. Routledge.
- [14] Noel J Entwistle and Dorothy Entwistle. 1970. The relationships between personality, study methods and academic performance. *British Journal of Educational Psychology* 40, 2 (1970), 132–143.
- [15] Mary Forehand. 2010. Bloom's Taxonomy: Original and Revised. *Emerging Perspectives on Learning, Teaching, and Technology* (2010).
- [16] Catherine A Gao, Frederick M Howard, Nikolay S Markov, Emma C Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T Pearson. 2022. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *BioRxiv* (2022), 2022–12.
- [17] Ahana Ghosh, Advait Sarkar, Siân Lindley, and Christian Poelitz. 2026. An Experimental Comparison of Cognitive Forcing Functions for Execution Plans in AI-Assisted Writing: Effects On Trust, Overreliance, and Perceived Critical Thinking. *arXiv preprint arXiv:2601.18033* (2026).
- [18] Barbara A Greene. 2015. Measuring cognitive engagement with self-report scales: Reflections from over 20 years of research. *Educational Psychologist* 50, 1 (2015), 14–30.
- [19] Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, and Luke Zettlemoyer. 2018. Mapping Language to Code in Programmatic Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 1643–1652. doi:10.18653/v1/D18-1192
- [20] Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.
- [21] David R. Krathwohl. 2002. A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice* 41, 4 (2002), 212–218. doi:10.1207/s1543042tip4104\_2
- [22] Felix Kriegelstein, Maik Beege, Günter Daniel Rey, Paul Ginns, Moritz Krell, and Sascha Schneider. 2022. A systematic meta-analysis of the reliability and validity of subjective cognitive load questionnaires in experimental multimedia learning research. *Educational Psychology Review* 34, 4 (2022), 2485–2541.
- [23] Emily Kuang, Minghao Li, Mingming Fan, and Kristen Shinohara. 2024. Enhancing UX Evaluation Through Collaboration with Conversational AI Assistants: Effects of Proactive Dialogue and Timing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 3, 16 pages. doi:10.1145/3613904.3642168
- [24] Kathryn Ann Lambe, Gary O'Reilly, Brendan D Kelly, and Sarah Curristan. 2016. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ quality & safety* 25, 10 (2016), 808–820.
- [25] Tze Wei Liew, Su-Mae Tan, Tak Jie Chan, Yang Tian, and Faizan Ahmad. 2025. Cognitive Benefits of Employing Multiple AI Voices as Specialist Virtual Tutors in a Multimedia Learning Environment. *Human Behavior and Emerging Technologies* 2025, 1 (2025), 8813532.
- [26] Tze Wei Liew, Su-Mae Tan, Wei Ming Pang, Mohammad Tariqul Islam Khan, and Si Na Kew. 2023. I am Alexa, your virtual tutor!: The effects of Amazon Alexa's text-to-speech voice enthusiasm in a multimedia learning environment. *Education and information technologies* 28, 2 (2023), 1455–1489.
- [27] Tze Wei Liew, Su-Mae Tan, Teck Ming Tan, and Si Na Kew. 2020. Does speaker's voice enthusiasm affect social cue, cognitive load and transfer in multimedia learning? *Information and Learning Sciences* 121, 3/4 (2020), 117–135.
- [28] Richard E Mayer and Roxana Moreno. 1998. A cognitive theory of multimedia learning: Implications for design principles. *Journal of educational psychology* 91, 2 (1998), 358–368.
- [29] Richard E Mayer and Roxana Moreno. 2003. Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist* 38, 1 (2003), 43–52.
- [30] Blessing Mbalaka. 2023. Epistemically violent biases in artificial intelligence design: the case of DALLE-2 and Starry AI. *Digital Transformation and Society* 2, 4 (2023), 376–402.
- [31] Tamara Munzner. 2025. Visualization analysis and design. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Courses*. 1–2.
- [32] Duygu Mutlu-Bayraktar, Veysel Cosgun, and Tugba Altan. 2019. Cognitive load in multimedia learning environments: A systematic review. *Computers & Education* 141 (2019), 103618.
- [33] Muhammad Naeem, Wilson Ozuem, Kerry Howell, and Silvia Ranfagni. 2023. A Step-by-Step Process of Thematic Analysis to Develop a Conceptual Model in Qualitative Research. *International Journal of Qualitative Methods* 22 (2023), 16094069231205789. arXiv:https://doi.org/10.1177/16094069231205789 doi:10.1177/16094069231205789
- [34] OpenAI. 2025. Codex. <https://openai.com/codex/>. Accessed: Dec 2025.
- [35] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 102 (Nov. 2019), 15 pages. doi:10.1145/3359204
- [36] R Pekrun. 2020. Commentary: Self-report is indispensable to assess students' learning. *Frontline Learning Research*, 8 (3), 185–193.
- [37] Dhruv Sabharwal, Robin Kabba, and Kajal Srivastava. 2023. Artificial intelligence (ai)-powered virtual assistants and their effect on human productivity and laziness: Study on students of delhi-ncr (india) & fujairah (uae). *Journal of Content, Community and Communication* 17, 9 (2023), 162–174.
- [38] Sascha Schneider, Maik Beege, Steve Nebel, Lenka Schnaubert, and Günter Daniel Rey. 2022. The cognitive-affective-social theory of learning in digital environments (CASTLE). *Educational psychology review* 34, 1 (2022), 1–38.
- [39] Roei Schuster, Congzheng Song, Eran Tromer, and Vitaly Shmatikov. 2021. You Autocomplete Me: Poisoning Vulnerabilities in Neural Code Completion. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 1559–1575. <https://www.usenix.org/conference/usenixsecurity21/presentation/schuster>
- [40] John Sweller, Jeroen JG Van Merriënboer, and Fred GWC Paas. 1998. Cognitive architecture and instructional design. *Educational psychology review* 10, 3 (1998), 251–296.
- [41] Hua Wang, Mark Chignell, and Mitsuru Ishizuka. 2006. Are two talking heads better than one? when should you use more than one agent in e-learning?. In *Proceedings of the 11th International Conference on Intelligent User Interfaces* (Sydney, Australia) (IUI '06). Association for Computing Machinery, New York, NY, USA, 366–368. doi:10.1145/1111449.1111539
- [42] Matthew O Ward, Georges Grinstein, and Daniel Keim. 2010. *Interactive data visualization: foundations, techniques, and applications*. AK Peters/CRC Press.
- [43] Colin Ware. 2019. *Information visualization: perception for design*. Morgan Kaufmann.
- [44] Peter C Wason and J St BT Evans. 1974. Dual processes in reasoning? *Cognition* 3, 2 (1974), 141–154.
- [45] Cody Watson, Michele Tufano, Kevin Moran, Gabriele Bavota, and Denys Poshyvanyk. 2020. On learning meaningful assert statements for unit test cases. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering (Seoul, South Korea) (ICSE '20)*. Association for Computing Machinery, New York, NY, USA, 1398–1409. doi:10.1145/3377811.3380429
- [46] Tao Xiao, Christoph Treude, Hideaki Hata, and Kenichi Matsumoto. 2024. DevGPT: Studying Developer-ChatGPT Conversations. In *Proceedings of the 21st International Conference on Mining Software Repositories* (Lisbon, Portugal) (MSR '24). Association for Computing Machinery, New York, NY, USA, 227–230. doi:10.1145/3643991.3648400

Participant	Response
P1	It generated my desired output
P2	I trust Cline
P3	It worked
P4	I think what Cline wrote was well-documented and readable enough, so I felt that changes were not necessary.

**Table 1: Participants' Responses to Q10: Why did you/did you not make any more adjustments (with the code)?**

Question	P1 Correct?	P2 Correct?	P3 Correct?	P4 Correct?
Q1 What was the name of the folder you were working on?	Yes	Yes	No	No
Q2 What was the name of the first file created?	Yes	Yes	No	Yes
Q3 How many functions/methods did the generated script have?	No	No	No	No

**Table 2: Participants' Responses to the Recall Questions. Their responses were cross-referenced with the directory they worked on to verify if their responses were correct.**

## 7 Appendices

### 7.1 Survey Questions

**Q1** (Single Choice) How many years of professional software engineering experience do you have?

- < 1 year
- 1-5 years
- 6-10 years
- > 10 years

**Q2** [Recall] What was the name of the folder/directory you were working on? (no need for the entire path, the folder name would be sufficient)

**Q3** [Recall] What was the name of the first file created?

**Q4** [Recall] How many functions/methods did the generated script have?

**Q5** [Understand](Single Choice) "I can reliably provide a concise summary for the first function in the generated script"

- Strongly agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Strongly disagree

**Q6** [Understand] (Single Choice) "I can reliably provide a concise summary for the last function in the generated script"

- Strongly agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Strongly disagree

**Q7** [Analyze] (Single Choice) "I can reliably determine the order of functions/methods that will be called in the script"

- Strongly agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Strongly disagree

**Q8** [Analyze] (Single Choice) The script is able to handle a case when there are no excel files in the working directory

- Yes
- No
- Unsure

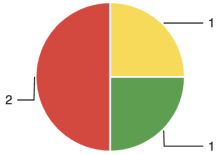
**Q9** [Evaluate] (Single Choice) Did you make some of your own manual adjustments with the script?

- Yes
- No

**Q10** [Evaluate] (Open Ended) Why did you/did you not make any more adjustments?

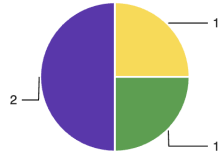
## 7.2 Survey Results

Q5 - "I can reliably provide a concise summary for the first function in the generated script"



Strongly disagree Somewhat disagree  
Strongly agree

Q6 - "I can reliably provide a concise summary for the last function in the generated script"



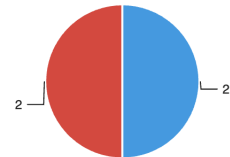
Strongly disagree Somewhat disagree  
Somewhat agree

Q7 - "I can reliably determine the order of functions/methods that will be called in the script"



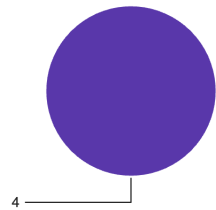
Strongly disagree Somewhat disagree  
Neither agree nor disagree Somewhat agree

Q8 - The script is able to handle a case when there are no excel files in the working directory



Unsure Yes

Q9 - Did you make some of your own manual adjustments with the script?



No