Constructive or Procedural? Student Engagement With an AI Tutor in Multivariable Calculus

Eva Politou
Harvard University

Megan Selbach-Allen
Harvard University

Large language models (LLMs) such as ChatGPT are increasingly used in higher education, raising questions about how they influence student learning. While unstructured use often leads to shallow engagement, carefully designed AI tutors may foster deeper reasoning. This study examines student engagement with an AI tutor in a multivariable calculus course, drawing on the ICAP framework. The focal task asked students to use dot products and vector reasoning to estimate storm velocity from radar data. Preliminary results suggest that while the AI tutor frequently posed prompts aligned with constructive engagement, student responses were most often active, reflecting a "completionist" approach to problem solving. Constructive reasoning was difficult to sustain, and tutor follow-up was sometimes limited. These findings highlight both the promise and the complexity of designing AI tutors to elicit and maintain constructive mathematical engagement.

Keywords: Large Language Models, AI Tutoring, Student Engagement, ICAP Framework, Multivariable Calculus

Literature Review

Large language models (LLMs) such as ChatGPT have quickly become part of higher education, with students and instructors experimenting with their use for studying, writing, and problem solving. Stakeholders describe LLMs as convenient, confidence-building, and accessible, but also express concerns about over-reliance, inequity of access, and diminished opportunities for peer or instructor interaction (Govender, 2023; Hasanein & Sobaih, 2023; Kasneci et al., 2023). Systematic reviews in mathematics education confirm that AI tools, especially chatbots, are increasingly common across secondary and higher education, though most work is exploratory or tool-focused rather than analyzing learning processes in detail (Awang et al., 2025; Almarashdi et al., 2024; Walkington, 2025).

Empirical studies present mixed evidence on the effectiveness of LLMs. On one hand, research suggests that unstructured use of ChatGPT may reduce cognitive engagement and lead to shallow learning. For example, MIT researchers found that students relying on ChatGPT for essay writing showed reduced neural engagement and poorer recall compared to those using search engines or working unaided (Kosmyna et al., 2024). Similar results emerged in inquiry-based tasks: students using ChatGPT reported lower cognitive load but also produced weaker justifications (Stadler et al., 2024). In physics education, students frequently accepted incorrect ChatGPT solutions uncritically and performed worse than peers using a search engine (Krupp et al., 2023). These findings underscore the risk that LLMs may encourage passive or procedural engagement unless structured to promote deeper reasoning.

On the other hand, carefully designed AI tutors show promise. In a recent randomized controlled trial, Kestin and colleagues (2025) demonstrated that students working with a custom AI tutor, TeachGPT, achieved significantly greater learning gains in less time than students in active-learning classrooms. Similarly, Bastani et al. (2024) found that while unstructured ChatGPT use could hinder later exam performance, a safeguarded "GPT Tutor" version, engineered to provide incremental hints and anticipate misconceptions, avoided these negative effects. A meta-analysis of 51 experimental studies similarly concluded that ChatGPT can substantially improve student performance and moderately support higher-order thinking, particularly when embedded in problem-based learning (Wang & Fan, 2025). In mathematics, ChatGPT-generated hints have produced learning gains comparable to those authored by human tutors, although accuracy concerns remain (Pardos & Bhandari, 2024). These findings suggest that the context and design of AI use are critical determinants of effectiveness.

One useful lens for understanding this contrast is tutoring research. Decades of work on human tutoring demonstrates that learning gains are often attributable not to tutor expertise but to the ways tutors elicit student thinking. Graesser et al. (1999) found that even untrained tutors often engaged students through feedback, hints, pumps, and elaborations, and that these discourse moves, not tutor expertise, were key features of effective tutoring. Chi et al. (2001) likewise found that tutoring effectiveness was driven by students' constructive responses, not tutor explanations. More recent innovations with AI tutors build on these insights, leveraging scaffolding strategies such as flipped dialogic approaches (Pavlova, 2024) or structured prompting frameworks (Tassoti, 2024) to push students beyond copying answers toward reflection and justification. Together, this body of work positions AI tutors not as replacements for human instructors but as systems that can be designed to elicit particular kinds of engagement.

Theoretical Framework

To analyze engagement systematically, this study draws on the ICAP framework (Chi, 2009; Chi & Wylie, 2014). ICAP proposes a hierarchy of engagement modes: Passive (receiving information), Active (manipulating or applying given information), Constructive (generating new inferences or explanations), and Interactive (co-constructing knowledge with peers). Research consistently shows that constructive and interactive activities lead to deeper learning than active or passive ones. In tutoring contexts, the constructive level is particularly salient: Chi et al. (2001) found that students' constructive responses predicted gains regardless of tutor expertise. Applications of ICAP to AI use likewise suggest that generative tools can promote deeper learning when designed to elicit constructive contributions (Gao et al., 2024).

We determined that in our context, interactive engagement is not feasible, since AI-tutor sessions involve a single student. We therefore adapt ICAP into three categories, Passive, Active, and Constructive, to code both AI prompts and student responses. Passive responses include minimal or off-task engagement (e.g., "idk," blank entries, or repeating AI text). Active responses recall facts, carry out straightforward operations, or follow procedures without interpretation. Constructive responses involve inference, justification, or proposing new ideas, such as setting up equations or reasoning about vector direction. We applied the same categories to AI prompts, distinguishing between explanatory (Passive), recall/procedural (Active), and generative/reflective (Constructive) tutor moves. This adaptation allows us to investigate how the design of AI prompts relates to the depth of student engagement.

Finally, while there is growing research on AI in mathematics education, most studies remain focused on procedural domains. ChatGPT explanations of multiplication, for instance, often reduce to repeated addition rather than deeper conceptual models (Hankeln, 2024). Studies on proof assistance emphasize correctness and formal revision rather than conceptual reasoning (Park & Manley, 2024). In calculus, student surveys report benefits of ChatGPT as a supplementary explainer but also concerns about accuracy and over-reliance (Serhan & Welcome, 2024). Reviews of AI in mathematics education confirm that most studies emphasize procedural problem solving, tool adoption, or broad perceptions (Almarashdi et al., 2024; Awang et al., 2025; Walkington, 2025). In contrast, very few investigate how AI tutors can foster conceptual reasoning in advanced mathematics.

This study fits in that gap by analyzing AI tutor logs from a multivariable calculus course to investigate how the phrasing of AI prompts shapes the depth of student engagement on a conceptual vector problem. Using an ICAP-based coding scheme, we classified both AI prompts and student responses as Passive, Active, or Constructive, allowing us to examine the relationship between tutor design and student engagement. In doing so, our work extends prior research that highlights both the promise and pitfalls of LLMs: when carefully designed, they can foster individualized learning and constructive reasoning, but in unstructured use they often default to procedural explanations and promote shallow engagement. At the same time, most studies of AI in mathematics education emphasize procedural tasks such as algebraic manipulation or proof revision, leaving open the question of how AI tutors might support conceptual reasoning in advanced courses.

Guided by this framework, we investigate two research questions:

1. What types of AI prompts lead to the most constructive student-generated responses?
2. To what extent do students demonstrate active or constructive engagement in AI-tutor sessions?

Methods

In this section, we describe the course context, participants, focal task, data collection procedures, and analytic approach.

Course Context

Data were collected during a multivariable calculus course offered at a selective private university in the US. The course enrolled approximately 400 students across multiple coordinated sections, each meeting three times per week with 25–30 students per section. In this course, “workshops” are weekly applied problem sets designed to highlight conceptual applications of multivariable calculus and real-world data. Traditionally, these workshops were completed in small groups with support from undergraduate course assistants. In Spring 2025, the course offered an AI tutor as an optional support, revised from an earlier Fall pilot to encourage shorter, more guiding prompts, delayed solution suggestions, and more frequent checks for understanding.

AI Tutor

The AI tutor platform used in this study, TeachGPT, was created by Dr. Greg Kestin at Harvard University. Prior to the course, the tutor was configured to be supportive and encouraging but not to provide direct answers. Instead, it was designed to elicit students’ thinking and ideas by offering prompts, hints, and checks for understanding.

Participants

This study was approved by the Institutional Research Board (IRB). Students were invited to participate via a survey distributed at the end of class, which asked about their experiences with the AI tutor and whether they consented to have their chat logs included in research. Of the total course enrollment, 123 students consented (approximately 31%).

Mathematical Task

The focal problem required students to use dot product calculations to build a system of equations for estimating the velocity of a storm using radar data from two different stations. Students worked with graphics and real data from a published research paper. Rather than providing exact numerical values, the task asked students to approximate storm speed and direction by interpreting vector arrows and color-coded visualizations. Evaluation emphasized conceptual application and reasoning processes, not exact numeric answers.

Data Collection

TeachGPT logs all student interactions, which can be downloaded by the instructor. The 123 consenting students produced over 1,500 turns of dialogue with the AI tutor. Each interaction turn (one uninterrupted chat entry) was recorded in the log. For this analysis, we focus on Spring 2025 logs, though the platform was also used in Fall 2024.

Data Analysis

Using the ICAP framework (Chi, 2009; Chi & Wylie, 2014), each turn of dialogue was coded as Passive, Active, or Constructive. Both AI prompts and student responses were coded, and turns could receive multiple codes. Two authors independently coded all responses and then met to reconcile disagreements. Agreement was generally high; precise interrater reliability will be reported in future analyses. At the time of this preliminary report, approximately 50% of the data has been coded. Codes are described in Table 1 below.

Table 1. Adapted ICAP-based coding scheme for student and AI tutor utterances.

Code	Student Responses	AI Prompts
P (Passive)	Minimal or off-task response; no meaningful engagement	Provides explanation without asking for a response
A (Active)	Recalls a fact, performs a basic operation, or follows a known procedure without interpretation	Prompts for recall or straightforward use of given information
C (Constructive)	Makes inferences, proposes ideas, sets up equations, estimates, or justifies choices	Prompts student to generate, infer, reflect, or justify

In what follows, we report descriptive patterns from the coded sample, focusing on the distribution of student and AI codes and illustrating how different types of AI prompts were associated with varying levels of student engagement.

Results and Discussion

Of the 564 turns coded so far roughly equally split between the student (n = 275) and the AI-tutor (n = 289) we found that the Active code occurred relatively equally between the students (n = 144) and the AI-tutor (n = 136). The passive code was actually more commonly applied to the AI tutor (n = 148) than the students (n = 98). However the form this code took was distinctly different. In the case of the students this code most often represented some amount of disengagement from the work as they said they just “didn’t understand” or asked the AI to give them the answer. The AI tutor got a passive tag when it provided substantial information to the student such as laying out a procedure for how to solve a portion of the problem. The passive code was often applied alone for the student but for the AI it often was coded alongside Active or Constructive. The largest difference was seen in the constructive code with the AI tutor (n = 146) receiving the code much more frequently than the student interactions (n = 35).

The preliminary analysis provides insight into the depth and character of student engagement in AI-tutor sessions. Constructive responses were relatively rare, with most turns coded as active and oriented toward carrying out procedures or recalling information. When constructive responses did occur, they were often followed by active turns rather than sustained elaboration, suggesting that constructive reasoning was more difficult to maintain over the course of an interaction. Many students appeared to approach the task with a "completionist mindset," focusing on moving efficiently through the problem rather than pausing to reflect or justify their ideas.

At the same time, the AI tutor frequently generated prompts that were well targeted to the mathematical content and designed to elicit reasoning. These prompts often invited inference, justification, or reflection, aligning with the constructive category of ICAP. However, constructive prompts did not always yield constructive responses. More often, students replied at the active level, providing correct but concise procedural steps rather than extended explanations.

Some challenges also emerged in the tutor's interactional pattern. The AI occasionally asked several questions in succession without fully engaging with student input, which may have limited opportunities for extended dialogue. In certain cases, it appeared to misinterpret student responses, resulting in exchanges that did not progress as productively as intended. These patterns would then constrain the tutor's ability to sustain constructive engagement, even when its questions were well designed.

Taken together, these findings suggest both the promise and the complexity of fostering constructive engagement through AI tutoring. The AI posed strong and targeted questions, yet students' responses did not consistently rise to the constructive level, and the tutor's responsiveness at times made it difficult to deepen reasoning. As a result, interactions often emphasized procedural progress more than conceptual exploration, highlighting opportunities to refine tutor design in ways that better support sustained constructive engagement.

Preliminary Nature of Report

Because this analysis is drawn from an ongoing project, the findings should be considered preliminary. We focus here on a single conceptual vector problem from one workshop, which generated over 1,500 turns of dialogue, of which 564 have been coded to date. This problem was selected because it emphasizes the kind of conceptual reasoning central to the course, provides a strong context for examining constructive engagement and we have film of students engaging with the problem during in-person workshops. As coding continues, our next steps include completing analysis of this problem and extending the comparison to video data from in-person workshop sessions, allowing us to contrast AI-facilitated and human-facilitated interactions. We invite feedback from the audience and engagement with the following question: what additional coding categories or analytic lenses could help capture engagement not fully addressed by ICAP?

References

Almarashdi, H. S., Jarrah, A. M., Khurma, O. A., & Gningue, S. M. (2024). Unveiling the potential: A systematic review of ChatGPT in transforming mathematics teaching and learning. *Eurasia Journal of Mathematics, Science and Technology Education*, *20*(12), em2555.

Awang, L. A., Yusop, F. D., & Danaee, M. (2025). Current practices and future direction of artificial intelligence in mathematics education: A systematic review. *International Electronic Journal of Mathematics Education*, *20*(2), em0823.

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2024). Generative AI without guardrails can harm learning. *The Wharton School Research Paper*.

Chi, M. T. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in cognitive science*, *1*(1), 73-105.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, *49*(4), 219-243.

Chi, M. T., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive science*, *25*(4), 471-533.

Gao, Z., Cheah, J. H., Lim, X. J., & Luo, X. (2024). Enhancing academic performance of business students using generative AI: An interactive-constructive-active-passive (ICAP) self-determination perspective. *The International Journal of Management Education*, *22*(2), 100958.

Govender, R. (2023). The impact of artificial intelligence and the future of ChatGPT for mathematics teaching and learning in schools and higher education. *Pythagoras*, *44*(1), 1-2.

Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & Tutoring Research Group. (1999). AutoTutor: A simulation of a human tutor. *Cognitive Systems Research*, *1*(1), 35-51.

Hasanein, A. M., & Sobaih, A. E. E. (2023). Drivers and consequences of ChatGPT use in higher education: Key stakeholder perspectives. *European journal of investigation in health, psychology and education*, *13*(11), 2599-2614.

Hankeln, C. (2024). Challenges in Using ChatGPT for Assessing Conceptual Understanding in Mathematics Education. *Journal of Mathematics Education*, *17*(1), 1-15.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, *103*, 102274.

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (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), 17458.

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X. H., Beresnitzky, A. V., Braunstein, I. & Maes, P. (2025). Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872*, *4*.

Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K. E., Lukowicz, P., Kuhn, J., ... & Karolus, J. (2023). Unreflected acceptance--investigating the negative consequences of ChatGPT-assisted problem solving in physics education. *arXiv preprint arXiv:2309.03087*.

Pardos, Z. A., & Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *Plos one*, *19*(5), e0304013.

Park, H., & Manley, E. D. (2024). Using ChatGPT as a proof assistant in a mathematics pathways course. *Mathematical Education*, *63*(2), 139-163.

Pavlova, N. H. (2024). Flipped dialogic learning method with ChatGPT: A case study. *International Electronic Journal of Mathematics Education*, *19*(1), em0764.

Serhan, D., & Welcome, N. (2024). Integrating ChatGPT in the Calculus Classroom: Student Perceptions. *International Journal of Technology in Education and Science*, *8*(2), 325-335.

Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior*, *160*, 108386.

Tassoti, S. (2024). Assessment of students use of generative artificial intelligence: Prompting strategies and prompt engineering in chemistry education. *Journal of Chemical Education*, *101*(6), 2475-2482.

Wang, J., & Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis. *Humanities and Social Sciences Communications*, *12*(1), 1-21.

Walkington, C. (2025). The implications of generative artificial intelligence for mathematics education. *School Science and Mathematics*.