

Hints Before Answers

Guardrails for an AI tutor, from the evidence on AI help and learning

The DuckyHelper team, MingLLM, Inc. October 2026 • Version 1.0 • hello@duckyhelper.com

Abstract

A 2025 field experiment found that high school students who practiced math with an unguarded GPT-4 chat did better during practice and worse on the exam once it was taken away, while a version with learning safeguards largely avoided the harm. Earlier research on tutoring systems and newer work on language model tutors point the same way: help that hands over answers can replace the thinking it was meant to support. This white paper describes the guardrails in DuckyHelper's tutor: teaching rules that separate questions of fact from problems to solve, a hint ladder that leaves the answer for last, feedback that marks where without fixing, and a memory of mistakes that brings them back. We are explicit about what the guardrails cannot do. DuckyHelper has not been evaluated in a controlled study. This paper reports no learning outcomes, no usage numbers and no benchmark scores of our own. Every number in it belongs to the study that reported it.

Keywords: generative AI, guardrails, hints, help seeking, AI tutoring, productive struggle

1 The evidence

1.1 An experiment in a high school

Bastani et al. (2025) gave high school math students one of two GPT-4 tutors during practice sessions. “GPT Base” was a plain chat. “GPT Tutor” was the same model with prompts designed to safeguard learning. Other students practiced without AI. During practice, grades rose 48% with GPT Base and 127% with GPT Tutor. On the exam, with no AI for anyone, the GPT Base group did 17% worse than students who never had access:

“unfettered access to GPT-4 can harm educational outcomes”

(Bastani et al., 2025)

Base and 127% for GPT Tutor). However, we additionally find that when access is subsequently taken away, students actually perform worse than those who never had access (17% reduction in grades for GPT Base)—i.e., **unfettered access to GPT-4 can harm educational outcomes**. These negative learning effects are largely mitigated by the safeguards in GPT Tutor. Without guardrails, students attempt to use GPT-4 as a “crutch” during practice problem sessions, and subsequently perform worse on

Figure 1. From the abstract of Bastani et al. (2025), PNAS, with the authors' own sentence highlighted. Reproduced for commentary.

The guarded tutor largely avoided this:

“These negative learning effects are largely mitigated by the safeguards in GPT Tutor”
(Bastani et al., 2025)

1.2 Why answers can hurt

The finding fits older research. In intelligent tutoring systems, students who used hints and feedback to dig out answers, rather than to learn, did much worse:

“students who frequently game the system score substantially lower on a post-test than students who never game the system”
(Baker et al., 2004)

Learners are not good judges of when to ask for help (Aleven et al., 2003), and they read effort as a sign of poor learning even when it is the opposite (Deslauriers et al., 2019). An assistant that removes all effort is pleasant to use and easy to over-use.

1.3 Language models default to answering

Recent work on LLM tutors finds the same tendency in the models themselves. Macina et al. (2023a) found that models good at solving math

“fail at tutoring because they generate factually incorrect feedback or are prone to revealing solutions to students too early”
(Macina et al., 2023a)

Sonkar et al. (2024) describe models that

“often provide immediate answers rather than guiding students through the problemsolving process”
(Sonkar et al., 2024)

and Dinucu-Jianu et al. (2025) argue that good teaching

“requires strategically withholding answers”
(Dinucu-Jianu et al., 2025)

When the AI is designed to teach, results look different. An AI tutor built on the course’s own teaching practices helped Harvard physics students learn more in less time (Kestin et al., 2025), and human tutors with AI suggestions were less likely to give the answer away and their students mastered more topics (Wang et al., 2024b).

2 The guardrails

DuckyHelper’s tutor has guardrails at three levels: the teaching rules it follows in every session, the tools it opens for the learner, and what it remembers between sessions.

2.1 Rules: facts get answers, problems get guidance

Refusing every answer would make the tutor useless for the many questions that are simply questions of fact. The rules draw the line by the kind of question:

IN DUCKYHELPER

A plain question of fact, or a request to explain an idea, gets a clear, short explanation, then a quick question to check that it landed.

IN DUCKYHELPER

On homework and graded work the tutor does not hand over the final answer. The learner tries first and is then guided one step at a time. A learner who is truly stuck works through a similar problem with the tutor, then goes back to their own.

2.2 The hint ladder

In the Mac app, a learner stuck on a problem gets a card with the problem and a ladder of hints they open one at a time: a nudge, a bigger hint, the worked step, and the answer last. The tool itself is defined so that no hint contains the answer, and the rules hold the last rung back:

IN DUCKYHELPER

The hint ladder is climbed one rung at a time, and the answer is held back until the learner asks for it a second time.

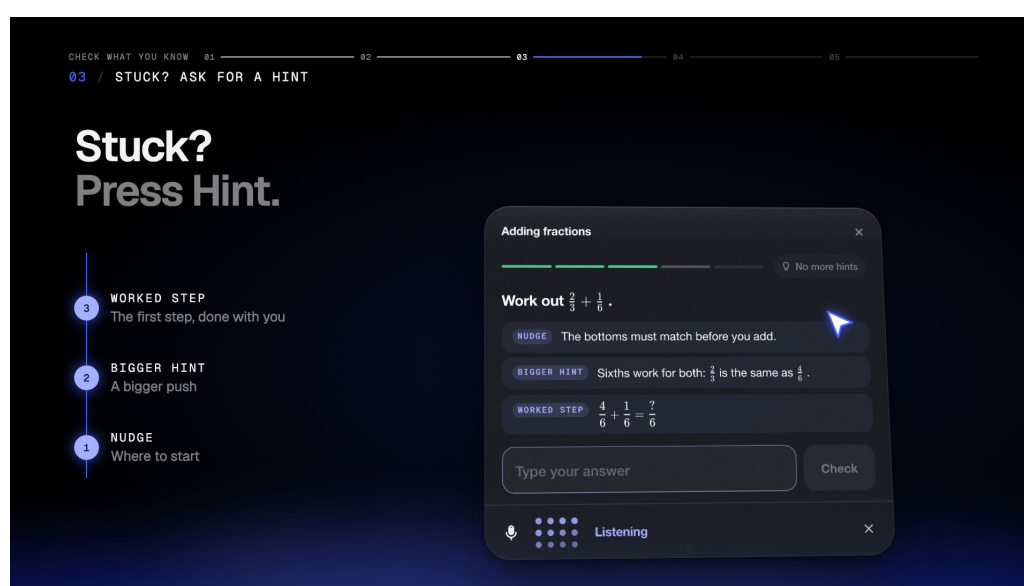


Figure 2. The hint ladder: Nudge, Bigger hint, Worked step, opened one at a time. Frame from our tutorial film, real app interface.

Hints used count against the tutor's estimate of the learner's mastery of that skill, so leaning on them keeps the skill in practice rather than marking it learned.

2.3 Quizzes and steps that hold back

IN DUCKYHELPER

While the learner works through a quiz, the tutor stays quiet. A wrong answer gets a one-sentence nudge, and a request for a hint gets a hint, not the answer.

IN DUCKYHELPER

On homework, worked steps stop where the learner is and add a hint for the next step, instead of showing the whole solution.

On Eddy's blackboard in the web app, math and homework are worked step by step and the answer is never simply said.

2.4 Feedback that marks where, not what

When the learner gets something wrong, the tutor points at the place rather than replacing their work: a circle on their screen in the Mac app, the first wrong line from the math check, a short note under their step on the whiteboard.

IN DUCKYHELPER

On the whiteboard, the tutor marks the learner's first wrong step with a short note and leaves the fix to the learner.

2.5 The tutor does not do it for them

The Mac app can click and type on the learner's screen, which is useful for demonstrations. The rules keep that power away from finishing the learner's work:

IN DUCKYHELPER

The tutor will not complete or turn in the learner's work for them, and it never pays for or deletes anything on their behalf. Every other tool is there so the learner acts.

2.6 Mistakes come back

A guardrail that only blocks is half a design. Missed quiz questions become flashcards, and when the tutor notices a kind of mistake it saves a clean practice question for it, never the learner's wrong work. Those cards return on a spaced schedule (Cepeda et al., 2006), so the learner retrieves the idea again later (Roediger & Karpicke, 2006).

3 Where the line is hard

- **Asking twice.** A learner who asks for the answer twice gets it. We chose this over refusing, because a tutor that never relents is easy to abandon for one that always answers. It also means a determined learner can get answers.
- **What counts as homework.** The tutor has to judge whether a question is a fact to know or a problem to solve. It will sometimes judge wrong.
- **Frustration.** Holding back can frustrate. Research on learners' feelings found that bore-

dom was more persistent and more harmful than frustration (Baker et al., 2010), but a tutor still has to keep the learner going; quick hints and short steps are how we try.

4 Limits

- DuckyHelper has not been evaluated in a controlled study. This paper reports no learning outcomes, no usage numbers and no benchmark scores of our own. Every number in it belongs to the study that reported it.
- The guardrails in Section 2 are instructions to a language model and interface rules. We have not measured how reliably the model follows them.
- The hint ladder and the math check are in the Mac app only.

The right test is the one Bastani et al. (2025) ran: compare learners with and without the tutor on what they can do *after* it is gone.

5 Conclusion

The evidence does not say AI is bad for learning. It says AI that does the learner's thinking is bad for learning, and AI built to teach can help. DuckyHelper's guardrails are our reading of that evidence: answer facts, guide on problems, hint before answering, mark where instead of fixing, and bring mistakes back.

References

- Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R. (2003). Help Seeking and Help Design in Interactive Learning Environments. *Review of Educational Research*, 73(3), 277-320. <https://doi.org/10.3102/00346543073003277>
- Baker, R. S. J. d., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4), 223-241. <https://doi.org/10.1016/j.ijhcs.2009.12.003>
- Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game the System". *CHI*. <https://doi.org/10.1145/985692.985741>
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), e2422633122. <https://doi.org/10.1073/pnas.2422633122>
- Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3), 354-380. <https://doi.org/10.1037/0033-2909.132.3.354>
- Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences*, 116(39), 19251-19257. <https://doi.org/10.1073/pnas.1821936116>
- Dinucu-Jianu, D., Macina, J., Daheim, N., Hakimi, I., Gurevych, I., & Sachan, M. (2025). From Problem-Solving to Teaching Problem-Solving: Aligning LLMs with Pedagogy using Reinforcement Learning. *EMNLP*. <https://doi.org/10.18653/v1/2025.emnlp-main.15>

- 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. <https://doi.org/10.1038/s41598-025-97652-6>
- Macina, J., Daheim, N., Chowdhury, S., Sinha, T., Kapur, M., Gurevych, I., & Sachan, M. (2023a). Math-Dial: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems. *Findings of EMNLP 2023*. <https://doi.org/10.18653/v1/2023.findings-emnlp.372>
- Roediger, H. L., & Karpicke, J. D. (2006). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science*, 17(3), 249-255. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>
- Sonkar, S., Ni, K., Chaudhary, S., & Baraniuk, R. (2024). Pedagogical Alignment of Large Language Models. *Findings of EMNLP*. <https://doi.org/10.18653/v1/2024.findings-emnlp.797>
- Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., & Demszky, D. (2024b). Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise. *arXiv preprint (v1)*. <https://arxiv.org/abs/2410.03017v1>

About this paper: written by the DuckyHelper team at MingLLM, Inc. It describes the design of DuckyHelper and cites published research. It has not been peer reviewed and reports no study of DuckyHelper. HTML version: <https://duckyhelper.com/research/papers/hints-before-answers/>