When Personalization Moves Latent Mastery but Not the Test: An LLM-Simulated Study of Knowledge Tracing and Study-Method Selection

Animesh Alang
University of California, Santa Cruz
aalang@ucsc.edu

Abstract

All learners in this paper are LLM-simulated; nothing here is evidence about real students. I ask a narrow question about AI tutors of the kind I am building in dino.study: if an LLM tutor keeps a per-student mastery estimate and picks a study method from that estimate, does anything measurable improve over a non-personalized LLM tutor? I run three arms (N=30 simulated students each) on a fixed introductory-probability item bank: (i) generic explain-then-practice, (ii) knowledge tracing with adaptive pacing, (iii) tracing plus method selection among retrieval practice, spaced repetition, worked examples, and Socratic questioning. On 8-item pre/post/delayed tests, normalized gains do not differ (all near zero; permutation p>0.25). On the simulator's latent mastery, they do: total mastery gain is 0.070 (control), 0.156 (trace), 0.229 (trace+method), all pairwise p<0.001. LLM mastery estimates track latent state (r=0.85) but next-item AUC is chance ($\approx$ 0.50); BKT is no better. A sticky method policy locks onto one method in 90% of students; an anti-lock-in fix removes lock-in without changing latent gains. Equalizing method gain constants and swapping in random/oracle methods barely moves outcomes. The usable claim is therefore modest and process-level: under this disclosed simulator, tracing changes latent mastery; short MC tests cannot see it; method selection adds a further latent bump that is partly built into the gain table.

1 Introduction

Bloom's two-sigma result (Bloom, 1984) still frames why people want tutors. LLM dialogue tutors are now easy to ship (Macina et al., 2023; Wang et al., 2024), but effects go both ways: structured AI tutoring can help (Kestin et al., 2025), and unguided GPT access can hurt (Bastani et al., 2025). So the design knobs matter. Two old knobs are knowledge tracing (Corbett and Anderson, 1995; Piech et al., 2015) and study-method choice—retrieval (Roediger and Karpicke, 2006), spacing (Cepeda et al., 2006), worked examples (Sweller and Cooper, 1985).

I care about this because I am building dino.study, an AI learning app meant to track what each student knows and switch how it teaches. Before claiming anything about real users, I wanted a falsifiable dry run. This paper is that dry run on **LLM-simulated students** (Markel et al., 2023), with every modeling choice disclosed and every number taken from logs—not adjusted.

Contribution. Under a disclosed simulator with literature-inspired method gains: (1) tracing and tracing+method raise latent mastery over a non-personalized tutor; (2) 8-item MC forms are too noisy to detect that (N≈350/arm would be needed for a 0.10 gain gap); (3) LLM mastery estimates correlate with latent state yet fail at next-item prediction; (4) naïve method selection locks in, and fixing lock-in does not buy more latent gain.

2 Method

Simulation, not students. No humans. Each simulated learner has a latent mastery vector over six intro-probability concepts, a learning rate, method affinities, one or two misconceptions, and a short persona used only to condition dialogue. Item answers are drawn from a 3PL-like model (guess 0.25, slip 0.08). Tutoring dialogue is generated by Claude (CLI); the model also returns engagement, scaffold quality, coverage, and a mastery estimate. After each episode, latent mastery increases by $\Delta = g(m) \cdot \text{affinity} \cdot \text{lr} \cdot \text{quality} \cdot \text{coverage} \cdot (1 - \theta)$, with fixed a-priori $g(m)$ (retrieval 0.18, spaced 0.16, Socratic 0.15, worked examples 0.14, generic explain 0.10). I did not retune $g(m)$ after seeing tables.

Arm	Gain	Delayed	Latent $\sum \Delta$	Ped. qual.	Lock-in	Entropy
Control	-0.14 $[-0.31, 0.03]$	0.49	0.070 [0.061, 0.080]	0.57	1.00	0.00
Trace	-0.06 $[-0.26, 0.15]$	0.41	0.156 [0.135, 0.178]	0.70	1.00	0.00
Trace+Meth.	-0.01 $[-0.16, 0.15]$	0.43	0.229 [0.206, 0.251]	0.74	0.90	0.06
TM v2	-0.11 $[-0.28, 0.06]$	0.49	0.224 [0.201, 0.248]	0.74	0.00	0.80

Table 1: Means with 95% CIs where shown. Gain = normalized pre/post. Latent $\sum \Delta$ sums mastery updates over four episodes. Lock-in = fraction of students using one method for all episodes. TM v2 = anti-lock-in policy, same population.

Items and procedure. Six concepts; 8 MC items each for pre, post, and delayed (isomorphic). Four tutoring episodes per student, then posttest, stylized forgetting, delayed form. Seed 42; N=30 per arm.

Arms. **Control:** fixed concept order, generic explain, no trace. **Trace:** LLM mastery estimates steer pacing toward weak concepts; method fixed to worked examples. **Trace+Method:** same tracing; method chosen from estimated mastery and affinity. A follow-up **Trace+Method v2** uses the same population with an anti-lock-in rule (penalize the previous method, more noise).

Ablations (replay). On Trace+Method students I replay learning updates across 5 response seeds with: random method, fixed retrieval, fixed worked examples, oracle affinity, equalized $g(m)$=0.14, BKT pacing without LLM estimates, and control-generic. Replays keep logged pedagogical quality; they change methods and seeds only.

Metrics. Normalized gain; delayed accuracy; sum of latent Δ; method entropy / lock-in rate; AUC for next-item prediction; ECE. Primary between-arm tests: 5000-shuffle permutation tests.

3 Results

Tests miss what the latent state shows. Table 1: normalized gains are compatible with zero in every arm. Permutation tests on gains: Trace vs Control p=0.55, Trace+Method vs Control p=0.26, Trace+Method vs Trace p=0.69. Latent $\sum \Delta$ tells a different story: Trace vs Control $+0.086$, Trace+Method vs Control $+0.158$, Trace+Method vs Trace $+0.072$, all $p<0.001$. Pedagogical quality from the LLM judge is also higher in personalized arms (0.74 vs 0.57, $p<0.001$). Multi-seed response replay (5 seeds) keeps the same qualitative pattern: test gains stay near zero with seed-to-seed SD $\approx$ 0.06–0.11.

Why the tests fail. Gain SD $\approx$ 0.47 on 8 items. A rough 80%-power check says detecting a 0.10 gain gap needs on the order of 350 students per arm. Oracle next-item AUC using true θ is only 0.61; guessing alone caps signal.

Mastery estimates look calibrated to θ, not to items. After episodes, LLM mastery estimates correlate with true θ at r=0.85 (MAE 0.10). Pooled next-item AUC is 0.495 (LLM) vs 0.504 (BKT). ECE: LLM 0.070, BKT 0.135. So the tracer is not useless as a state estimator inside the simulator, but it is useless—here—as a next-item ranker.

Method selection mostly did not select. Trace+Method locked onto a single method for 90% of students (entropy 0.06). v2 drives lock-in to 0% (entropy 0.80) with no latent-gain change (0.224 vs 0.229, p=0.79). Replay ablations on Trace+Method students (mean gain across 5 seeds): observed -0.059, random method -0.110, fixed retrieval -0.060, oracle affinity -0.056, equalized $g(m)$ -0.064. Random is a bit worse; oracle is not better than the sticky policy on tests. Equalizing $g(m)$ barely moves replayed test gains, which is a warning: the latent Trace+Method bump partly rides on unequal $g(m)$ plus higher logged quality, not on clever switching.

4 Limitations

This is a simulator study. The learning update encodes method differences I chose from the literature; that is a feature for transparency and a bug for causal claims about methods. Dialogue quality is judged by the same model family that writes the dialogue. "Delayed retention" is a stylized decay, not a week later. The item bank is tiny. dino.study product metrics and any visa-related evidence need real learners and IRB oversight—not these numbers. I am an undergraduate with no prior publications; errors are mine.

5 Conclusion

In this LLM-simulated sandbox, personalization changes the latent mastery the sandbox tracks, and short multiple-choice tests do not show it. That is useful as a design warning for dino.study and as a methods warning for anyone evaluating tutors with LLM students alone. Code, prompts, raw JSON, and ablations are in the accompanying release.

References

Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. 2025. Generative AI without guardrails can harm learning. *Proceedings of the National Academy of Sciences*, 122(26):e2422633122.

Benjamin S. Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6):4–16.

Nicholas J. Cepeda, Harold Pashler, Edward Vul, John T. Wixted, and Doug Rohrer. 2006. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3):354–380.

Albert T. Corbett and John R. Anderson. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278.

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.

Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621.

Julia M. Markel, Steven G. Opferman, James A. Landay, and Chris Piech. 2023. GPTeach: Interactive TA training with GPT-based students. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*.

Chris Piech, Jonathan Spencer, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. In *Advances in Neural Information Processing Systems*. ArXiv:1506.05908.

Henry L. Roediger and Jeffrey D. Karpicke. 2006. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3):249–255.

John Sweller and Graham A. Cooper. 1985. The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, 2(1):59–89.

Rose E. Wang, Ana T. Ribeiro, Carly D. Robinson, Susanna Loeb, and Dorottya Demszky. 2024. Tutor CoPilot: A human-AI approach for scaling real-time expertise.