Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1552

Making AI Tutoring Productive: Evidence from a Mastery-Based Math Practice Experiment

Philip Oreopoulos
University of Toronto

Michael Liut
University of Toronto

Alp Sungu
University of Pennsylvania

Nina Low
University of Toronto

Large language models (LLMs) can make tutoring more scalable, but only if students use them to reason through mistakes rather than avoid effort. We study this question in a randomized field experiment with more than 6,000 middle-school students in Hamilton County Schools using NUMI, a research-based computer-assisted learning (CAL) platform. Students were randomized to AI versus CAL-only support, mastery versus non-mastery progression, and one of two math topics. One week later, they completed a delayed assessment covering practiced and unpracticed material. AI students progressed more slowly and attempted fewer questions, but answered more accurately conditional on reaching an attempt. The clearest mechanism appeared after mistakes: AI improved next-attempt correctness and reduced attempts needed to return to a correct answer, while increasing time spent on each question receiving structured support. Mastery increased three-correct-in-a-row attainment but did not by itself improve delayed learning. The most encouraging delayed-test evidence appears when AI is embedded in the mastery workflow, with marginally significant gains concentrated on practiced Exercise 1 material. The results suggest that LLM tutoring can add value over standard CAL, but its value depends on structure that turns mistakes into productive learning moments.

VERSION: August 2026

Suggested citation: Oreopoulos, Philip, Michael Liut, Alp Sungu, and Nina Low. (2026). Making AI Tutoring Productive: Evidence from a Mastery-Based Math Practice Experiment. (EdWorkingPaper: 26-1552). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/01qv-6c22

Making AI Tutoring Productive: Evidence from a Mastery-Based Math Practice Experiment

Philip Oreopoulos* Michael Liut† Alp Sungu‡ Nina Low§

August 17, 2026

Abstract

Large language models (LLMs) can make tutoring more scalable, but only if students use them to reason through mistakes rather than avoid effort. We study this question in a randomized field experiment with more than 6,000 middle-school students in Hamilton County Schools using NUMI, a research-based computer-assisted learning (CAL) platform. Students were randomized to AI versus CAL-only support, mastery versus non-mastery progression, and one of two math topics. One week later, they completed a delayed assessment covering practiced and unpracticed material. AI students progressed more slowly and attempted fewer questions, but answered more accurately conditional on reaching an attempt. The clearest mechanism appeared after mistakes: AI improved next-attempt correctness and reduced attempts needed to return to a correct answer, while increasing time spent on each question receiving structured support. Mastery increased three-correct-in-a-row attainment but did not by itself improve delayed learning. The most encouraging delayed-test evidence appears when AI is embedded in the mastery workflow, with marginally significant gains concentrated on practiced Exercise 1 material. The results suggest that LLM tutoring can add value over standard CAL, but its value depends on structure that turns mistakes into productive learning moments.

*University of Toronto and NBER. Email: philip.oreopoulos@utoronto.ca.
†University of Toronto. Email: michael.liut@utoronto.ca.
‡The Wharton School, University of Pennsylvania. Email: alpsungu@wharton.upenn.edu.
§University of Toronto. Email: z.low@utoronto.ca.

Keywords: artificial intelligence, tutoring, computer-assisted learning, mastery learning, mathematics education, randomized experiment

JEL Codes: I21, I24, I28, C93, O33

Acknowledgments: We are grateful to Hamilton County Schools, participating teachers, school leaders, and students for making this study possible. We thank Kassidy Cunningham and John Rice, our site partners at Hamilton County School District, for their collaboration and support throughout this project. We are also grateful to Aaditya Dhingra, Matthew Wittman, Akhil Choraria, Yuvaansh Kapila, Thanh Nguyen, and Najwa Ibrahimi for developing the NUMI platform, and to Dellannia Segreti and Ruochong Dong for excellent research assistance. This research was supported by the Smith Richardson Foundation, the Social Sciences and Humanities Research Council of Canada (SSHRC), the National Sciences and Engineering Research Council of Canada (NSERC), and the Abdul Latif Jameel Poverty Action Lab (J-PAL). This study was registered with the AEA RCT Registry under trial number AEARCTR-0018678. The views expressed here are those of the authors and do not necessarily reflect those of Hamilton County Schools, J-PAL, the Smith Richardson Foundation, the Social Sciences and Humanities Research Council of Canada, or any other participating organization. All errors are our own.

1 Introduction

Tutoring works in part because a good tutor responds when a student gets stuck. The tutor can notice a misconception, ask smaller questions through natural scaffolding, encourage another attempt, or slow the student down before a mistake becomes frustrating. These features are difficult to reproduce at scale. Computer-assisted learning platforms can provide videos, practice questions, feedback, and worked solutions to many students at low cost, but they are usually less able to respond flexibly to what a student is thinking in the moment. Large language models (LLMs) create the possibility of combining the scale of computer-assisted learning with some of the personalized responsiveness of tutoring.

That possibility is not enough because the technology does not determine how it is used. A student who has access to an AI tutor may not be motivated to use it, or want to use it just to ask for the answer, or click through hints or explanations without much reflection. These are not small implementation details. They are choices central to whether AI tutoring improves or harms learning. The same technology that can explain a misconception can also reduce the need for the student to think through the problem independently. Which of these two outcomes occurs depends not only on the model's capabilities but also on the pedagogical design of the learning environment and the behavior it elicits from students.

The question is therefore behavioral as much as technological: not only whether an AI tutor can explain well, but also whether the learning environment leads students to use those explanations to think. We study this with NUMI, a computer-assisted learning platform built specifically to evaluate AI tutoring in middle-school math. NUMI combines short instructional videos, practice problems, immediate feedback, worked solutions, and a guard-railed AI tutor embedded directly beside the math exercise. The AI tutor was designed to avoid giving direct answers and instead provide structured support. Its features include many nudges designed to shape student behavior towards engaging with it more (see Online Appendix Section A for details).

The experiment was conducted in middle-school math classes in Hamilton County Schools, which includes Chattanooga, Tennessee. Teachers chose one of their math class periods during one week in the end of March, 2026 for students to complete a NUMI assignment, and administered a short delayed assessment during another class approximately one week later. The activity was presented as a review practice for material relevant to the upcoming state assessment.

Students were individually randomized at login along three dimensions. First, they were assigned to practice one of two grade-specific topic bundles. Second, they were assigned either

to receive access to NUMI's AI tutor or to a CAL-only (no AI) version of the platform. Third, they were assigned either to a mastery version or a non-mastery version of the activity. In the mastery condition, students had to watch at least one minute of the relevant video and answer three exercise questions in a row correctly before moving to the next part of the assignment. In the non-mastery condition, students could move more freely through the content after attempting at least one question. The delayed assessment, one week later, contained one question corresponding to each exercise type for each topic bundle, so every student had both practiced-topic and unpracticed-topic outcomes.

The design was motivated by the hypothesis that AI tutoring may be more effective when students have a reason to use it. Without a mastery requirement, a student who makes a mistake can view a solution, skip ahead, or move on. With mastery, a student must answer three questions in a row correctly, so using the AI as a crutch only leaves them stuck. The next question still needs to be solved independently. The rule thus gives students a reason to use the AI to understand the step rather than skip past it. The mastery rule may therefore create demand for help. At the same time, it may slow students down and reduce exposure to later material.

We find that both parts of this hypothesis matter. Mastery worked as a behavioral lever. It increased the amount of practice students completed, the number of correct answers during practice, and increased the probability of reaching three correct answers in a row. In the initial exercise given to students during a a math session, mastery increased the number of questions attempted and completed by 4.62, the number of questions answered correctly by 1.23, and the probability of reaching three correct answers in a row by 28.7 percentage points. Yet mastery alone did not generate clear gains on delayed-test performance.

AI changed practice in a different way — not by how much students did, but how they did it. Students assigned to AI attempted fewer questions and moved more slowly. In the first exercise, AI reduced the number of questions completed by about one question and increased time spent working on a question by about 1.64 minutes. In the mastery condition, AI slowed progress further, reducing the number of Exercise 1 questions completed relative to mastery without AI. However, this slower pace came with higher accuracy for the questions that students actually worked on. The platform data therefore point to a pace-quality tradeoff: AI does not help students finish more work, it changes the quality and timing of the work they complete.

The clearest evidence of this change appears after mistakes. In a stacked analysis of mistake events, AI increased the probability that the next attempt was correct and reduced the number of attempts needed to reach the next correct answer. These effects were larger

in the mastery condition. At the same time, AI increased the time it took to get to the next correct answer, consistent with the fact that students were being guided through a structured walkthrough rather than immediately guessing another answer. This is the mechanism at the center of the paper. AI appears to make recovery from mistakes more productive, but not faster.

The one-week-later test results are modest, but positive. The strongest and cleanest learning signal comes from practiced Exercise 1 material among mastery students. AI students in the mastery condition score about three percentage points higher than CAL-only students on the practiced Exercise 1 delayed-test question, while showing essentially no difference on the corresponding unpracticed Exercise 1 question. The AI versus no-AI estimate is only marginally statistically significant, but roughly the same size as the practice effect itself. These results do not support a claim that a short AI-supported practice session produces transformational learning gains. However, they provide evidence that an LLM-based tutor can add value over a comparable CAL-only platform, especially when embedded in a structure that makes post-mistake support salient and useful.

The paper contributes to a rapidly growing literature on generative AI in education. Existing evidence is mixed. Some studies show that AI-generated explanations or AI support for human tutors can improve learning or instructional quality. Other studies show that open-ended AI access can improve performance while students have the tool but harm later unaided performance. The closest comparison is Bastani et al. (2025), who find that unrestricted GPT-4 access improves high-school math practice performance but reduces subsequent unaided test performance; a guard-railed GPT tutor mitigates this harm but does not improve learning relative to the no-AI condition. NUMI differs by embedding a guard-railed AI tutor inside a purpose-built CAL platform and testing its value-added relative to the same platform without AI. The goal is not to evaluate generic generative AI access, but to ask whether a proactive AI tutor can improve learning when placed inside a structured classroom practice environment.

The paper also contributes to the literature on computer-assisted learning and mastery learning. CAL platforms can scale practice and feedback, but their effects depend heavily on student engagement, teacher implementation, and whether students use the platform in ways that promote learning. Mastery rules are often intended to ensure that students practice until they understand the material. The results here suggest a more cautious interpretation. The mastery rule increased apparent mastery during practice, but did not by itself produce clear delayed gains. Structure may create learning opportunities, but the quality of support during those opportunities matters.

2 Related Literature and Conceptual Framework

This paper relates to three areas of literature: research on tutoring and computer-assisted learning, research on mastery learning and structured practice, and a rapidly growing literature on generative AI in education. The common theme across these is that learning depends not only on access to instructional resources, but also on how students engage with them.

2.1 Tutoring, computer-assisted learning, and the challenge of scale

Tutoring is one of the most effective educational interventions studied. Recent reviews and meta-analyses find large average effects from tutoring, especially when tutoring is frequent, curriculum-aligned, and delivered by trained tutors (Nickow et al., 2020; Dietrichson et al., 2017; Ritter et al., 2009). A tutor can adapt to a student's level, respond to confusion, provide encouragement, and decide when to give a hint, ask a question, or move on. The problem is scale. High-quality tutoring is often costly, requires human time, training, scheduling, and sustained student participation, making it difficult to provide to all students who might benefit.

Computer-assisted learning offers a different tradeoff. It can deliver instructional videos, practice questions, feedback, and worked examples at low marginal cost. Evidence from educational technology interventions is mixed, with impacts depending heavily on the quality of the software, implementation, and how instructional time is used (Banerjee et al., 2007; Barrow et al., 2009; Bulman and Fairlie, 2016; Escueta et al., 2020). More recent work shows that well-designed adaptive learning platforms can generate meaningful gains, especially when they personalize content to students' current level (Muralidharan et al., 2019; Ganimian et al., 2023). But CAL platforms also face a central implementation problem: students must actually engage with the material in ways that produce learning. They may rush, guess, skip explanations, or stop after minimal effort. Even when feedback is available, it may not be used at the moment when it would help most.

LLM-based tutoring is promising because it may relax some of the constraints faced by both tutoring and CAL. Compared with traditional CAL, an LLM tutor can respond in natural language, ask follow-up questions, and tailor explanations to a student's apparent misconception. Compared with human tutoring, it can be offered to many students at once at low marginal cost. But this promise also creates a risk. A model designed to be helpful may be too helpful, allowing students to obtain answers without doing the work that produces learning. This risk is closely related to longstanding concerns in learning science about passive receipt of worked solutions, reduced cognitive effort, and overreliance

on external support (Chi, 2000; Renkl, 2002; Bjork and Bjork, 2011; Koedinger et al., 2012). The relevant comparison is therefore not simply whether AI can answer math questions, but whether an AI tutor can improve learning when embedded in a real classroom practice environment.

2.2 Generative AI and learning

The early evidence on generative AI in education is mixed. Some studies show that AI-generated explanations or hints can support learning. Pardos and Bhandari (2024), for example, compare AI-generated help with human tutor-authored help in an online text-to-text tutoring session in class. They find that AI-generated help produces learning gains comparable to human-authored help, while also documenting important quality-control problems: a human was necessary in this setup to check and agree on each AI-suggested response. This suggests that LLMs can generate useful instructional content, but also that educational deployment requires design and monitoring. Other studies find positive effects of AI-supported learning in particular contexts, including AI-assisted homework in language learning (Vanzo et al., 2024), AI tutoring in college physics (Kestin et al., 2025), and supplemental AI math tutoring in Ghana (Henkel et al., 2024). These studies point to the potential for AI to expand access to interactive support, but they differ from our setting in treatment comparison, age group, subject matter, and whether AI is added to an otherwise similar CAL environment.

At the same time, other evidence warns that AI access can improve performance while students have the tool without improving, and sometimes even reducing, later unaided performance. The closest study to ours is Bastani et al. (2025), who evaluate GPT-4-based tutors in high-school mathematics. Students were randomly assigned to no AI, an open-ended GPT-4 interface, or a guard-railed GPT tutor. GPT access improved practice performance while students had the tool, especially in the guarded tutor condition. But open-ended GPT access reduced subsequent unaided exam performance, suggesting that students used the tool in ways that substituted for learning. The guarded tutor mitigated this harm, but did not generate clear positive learning gains relative to the no-AI condition. Why the guard-railed tutor produced no learning gain is unclear; one possibility is that students disengaged once they realized it would not give them the answers.

Evidence from programming and other domains reaches a similar conclusion. Lehmann et al. (2024) study LLM use in coding classes and distinguish between students who use LLMs as tutors and those who rely on LLMs to solve practice exercises for them. They find that conversational, explanation-seeking use of AI is correlated with positive learning gains, while solution-seeking use is not. Related work with online courses and programming

support also emphasizes that AI may reduce engagement with course materials or change the way students allocate effort (Nie et al., 2024). These findings are consistent with a broader concern: AI can make students feel supported and productive while reducing the amount of independent reasoning they do.

In computing education research, LLM tutors have been developed with explicit guardrails that withhold direct solutions. Liffiton et al. (2023), for instance, introduce CodeHelp, an always-available LLM assistant that scaffolds programming students toward answers without revealing them, and Kumar et al. (2023) present QuickTA, an instructor-configurable conversational agent prompted to help students brainstorm solutions without supplying them directly. Related work uses LLMs to generate practice problems, improve error messages, and classify the kind of help students request (Sarsa et al., 2022; Leinonen et al., 2023; Savelka et al., 2023), and Kumar et al. (2024) shows how students are guided to interact with an LLM, not merely whether they have access, shapes performance, trust, and query behavior. This work shares NUMI's central design commitment that the tutor should guide reasoning rather than just supply answers, but it has largely been evaluated through usage patterns and student perceptions within single courses, rather than through randomized comparisons of delayed learning against an otherwise identical platform without AI.

Other work studies AI as a support for human instructors rather than as a direct tutor for students. Wang et al. (2025) evaluate Tutor CoPilot, a tool that provides real-time suggestions to human tutors. In a randomized trial involving K–12 tutoring, students assigned to tutors with AI support were more likely to master math topics, with especially large gains among initially lower-rated tutors. The mechanism appears to run through changes in tutor behavior: AI-supported tutors were more likely to ask guiding questions and less likely to give away answers. This evidence is important because it shows that AI can improve learning by improving the quality of instructional interactions. But it leaves open a different question: can AI add value when it interacts directly with students inside a CAL platform, without a human tutor mediating each response?

Recent reviews reach similar conclusions. The Stanford AI Hub review of AI in K–12 education finds rapid growth in the number of AI-related education studies, but a much smaller number of rigorous causal evaluations (Fesler et al., 2026). It summarizes the current evidence as showing that AI often improves performance while students have access to the tool, but that evidence on independent learning, transfer, and longer-run outcomes remains limited and mixed. The review also emphasizes that pedagogical design matters: tools that scaffold reasoning and discourage answer-copying appear more promising than general-purpose AI access.

Our study contributes to this literature in three ways. First, it evaluates an LLM tutor in a large, real classroom setting with middle-school students. Second, it compares AI support to an otherwise similar CAL-only environment, rather than to no technology, standard instruction, or extra instructional time. Third, it tests not only whether AI access matters, but whether its effects depend on structure that gives students a reason to use the AI productively.

2.3 Mastery learning, structured practice, and productive engagement

The experiment also relates to research on mastery learning and structured practice. Mastery learning starts from a simple idea: students should keep working until they demonstrate understanding. This idea originates with Bloom (1968), who argued that most students can reach high achievement if instruction lets them master each unit before progressing, and later work has synthesized its classroom implementation (Guskey, 2010). Meta-analytic evidence suggests mastery-based programs can raise achievement on average (Kulik et al., 1990), though effects depend heavily on how mastery is operationalized. In computer science, mastery approaches have been applied but the evidence remains sparse and without generalizable results (Garner et al., 2019).

In practice, platforms often implement this idea with progression rules, such as requiring a student to answer several questions correctly before moving on. Such rules can increase practice and prevent students from skipping ahead too quickly. But a platform mastery rule is not the same thing as mastery. A short-run streak of correct answers may reflect understanding, but it may also reflect persistence, item familiarity, lucky guesses, or repeated exposure to similar problems. Generative AI sharpens this concern: a recent Dagstuhl seminar warned that AI can produce a false sense of mastery" or the "illusion of learning", in which platform-defined progress outpaces durable understanding, especially for struggling students (Falkner et al., 2024). Participants noted that AI is most likely to function as a crutch precisely for struggling students, and that mastery rules therefore depend heavily on whether the support students receive after errors promotes reasoning rather than answer-getting. Our design speaks directly to this question by varying whether mastery is paired with a guard-railed AI tutor.

This distinction is especially important in a fixed class period. Requiring students to continue until they reach a threshold can increase practice on early material while reducing time available for later material. It can also change which students reach later exercises. A

mastery rule may therefore improve learning, but it may also create exposure and selection complications.

This issue connects to a broader literature on productive struggle, effort, and feedback. Learning often improves when students actively retrieve information, explain reasoning, and work through errors, rather than passively observe solutions (Roediger and Karpicke, 2006; Dunlosky et al., 2013; Koedinger et al., 2012). At the same time, struggle is productive only when students receive enough support to avoid frustration, disengagement, or repeated uninformative attempts (Vygotsky, 1978; Wood et al., 1976; Kapur, 2008). The educational value of a mastery rule therefore depends on what happens after a student makes a mistake. If the rule simply requires more attempts, it may increase persistence without increasing understanding. If it pairs persistence with targeted feedback and reflection, it may improve learning.

In NUMI, the mastery condition required students to answer three exercise questions in a row correctly before progressing. This rule was designed partly as a learning intervention and partly as a way to create stronger incentives to engage with support. If students can move on after a mistake, they may have little reason to ask for help or study the explanation carefully. If they must keep working until they reach a threshold, then mistakes become more consequential. In that environment, an AI tutor may be more useful because students have a reason to understand what went wrong.

This logic suggests complementarity between structure and AI. Structure creates moments of need. Students must persist after mistakes. AI can make those moments more productive by diagnosing likely misconceptions, asking smaller questions, and guiding students through the next step. But structure alone may not be enough. If students respond to a progression rule by repeatedly trying answers until they reach a streak, the rule may increase platform-defined success without increasing durable learning.

3 Intervention, Experimental Design, and Data

3.1 Setting

The experiment took place in Hamilton County Schools, which includes the city of Chattanooga, Tennessee. The study was conducted during two fixed school weeks in March and early April. During the first week, March 23–27, 2026, teachers selected one regular math class period for students to complete the NUMI practice activity. The activity was designed to last approximately 50 minutes within a standard class period. During the following week,

March 30–April 3, teachers selected a class period to administer a short delayed assessment, designed to take approximately 15–20 minutes.

The study included students in grades 6 through 8 across 20 schools and just under 100 teachers. Participation was universal and advertised as part of a small review for an upcoming state assessment. To reduce variation in implementation, we provided students with headphones, paper, and pencils. Teachers were asked to introduce the activity using a common script. Students were told that the activity included videos, practice exercises, an exit-ticket quiz, and a corresponding test the following week. They were instructed to work silently, use paper and pencil, think carefully, and not rush or guess. Teachers were instructed to help with technical issues but not with math questions.

The delayed assessment the following week was also administered in class using a common script. Students were told that the assessment covered material from the previous week's NUMI activity and were asked to work silently, use paper and pencil, and do their best. Teachers again were instructed not to answer content questions.

3.2 The NUMI platform

NUMI is a purpose-built, AI-powered computer-assisted learning platform for middle-school math practice. The platform was designed both as a learning tool and as an experimental environment. It delivered common instructional content, randomized students across treatment arms, recorded detailed practice behavior, and administered delayed assessments. NUMI was built for this project by a development team composed primarily of current and recent undergraduate and graduate computer-science students from the University of Toronto.

Each NUMI assignment began with an instructional video on a grade-specific math topic, followed by practice exercises and an exit ticket. Students were randomly assigned to one of two topic bundles for their grade. Each topic bundle contained two exercises. Exercise 1 introduced practice problems closely tied to the video. Exercise 2 built on the same topic and was generally more challenging. The delayed assessment contained one question corresponding to each exercise for each topic bundle. As a result, each student received two delayed-test questions on material they had practiced and two delayed-test questions on material from the other topic bundle that they had not practiced in NUMI.

The grade-level topics were chosen to align with middle-school math review content relevant to the state assessment. For grade 6, the topics included fraction division and percent problems. For grade 7, the topics included proportional reasoning and operations

with rational numbers. For grade 8, the topics included equations with variables on both sides and expressions involving powers or square roots. The practice and delayed-test questions were generally of the same type, although students had not seen the delayed-test questions before. Online Appendix section A shows examples of the student interface and practice questions.

3.3 CAL-only and AI support

All students received a computer-assisted learning environment with videos, practice questions, immediate feedback, and worked solutions. Students assigned to the CAL-only condition did not have access to AI-generated hints, chat, or walkthroughs. When they submitted an incorrect answer, the platform displayed the worked solution. Students could use the solution to review the problem before continuing.

Students assigned to CAL-only saw the same instructional videos, worked on the same practice items, and had access to worked solutions. The key difference between the AI and CAL-only conditions is the presence of the AI tutor: AI students could receive first-step hints, interactive walkthroughs after mistakes, and step-specific explanations generated by NUMI. The main AI comparison therefore estimates the value added of an LLM-based tutor over a standard CAL platform that only displays solutions after making mistakes.

NUMI's AI functionality provided support in several ways. First, students could occasionally click a "Help me get started" button before attempting a question. The button was designed to nudge students towards working through the problem by getting them to think about where to begin. NUMI introduced the approach needed for the first step, but the key hint was delivered as a short multiple-choice reasoning question with two options. One option reflected the correct first step and the other reflected a plausible misconception. The student therefore still had to think about what to do next. After the student responded, NUMI stated whether the response was correct. If correct, NUMI briefly explained why and encouraged the student to work through the rest of the problem independently. If incorrect, NUMI identified the correct reasoning and explained why the alternative was wrong before sending the student back to the problem. This help was not continuously available; after use, the help button reappear only until the student had attempted two additional questions.

3.4 Mastery and non-mastery progression

Students were also randomly assigned to either mastery or non-mastery progression.

In the mastery condition, students were required to spend at least one minute on the instructional video before they could skip it. The platform encouraged them to watch the video because it would help with the exercise. During practice, mastery students had to answer three questions in a row correctly before progressing to the next part of the assignment. If they made a mistake, the streak resets. This rule applied regardless of AI assignment.

In the non-mastery condition, students had more flexibility. They could skip the video at any time. During practice, after attempting three questions, the platform told them how many they had answered correctly and asked whether they wanted to continue practicing or move on. Non-mastery students were encouraged to study until they understood the material, but were not required to reach three correct answers in a row before progressing.

The mastery condition was intended to serve two purposes. First, it was a learning intervention: requiring students to keep practicing until they demonstrated short-run success might increase understanding. Second, it was a way to create stronger incentives to engage with support. If students must continue until they reach a threshold, then mistakes become more consequential. Therefore, one of our pre-registered hypotheses is that mastery and AI support might be complements. At the same time, the design also creates a potential exposure tradeoff: students who spend more time reaching mastery on Exercise 1 may have less time to reach Exercise 2 or the exit ticket.

3.5 Randomization

Students were randomized individually at initial registration. Randomization occurred independently across three dimensions: topic assignment, AI assignment, and mastery assignment. Each student had an equal chance of being assigned to one of the two grade-specific topic bundles, to AI or CAL-only support, and to mastery or non-mastery progression. Teachers and students did not know assignments in advance.

The independent randomization created a $2 \times 2 \times 2$ design: topic A versus topic B, AI versus CAL-only support, and mastery versus non-mastery progression. Much of the analysis focuses on the 2×2 design crossing AI and mastery, pooling across topic assignment while using the topic randomization to define practiced and unpracticed delayed-test outcomes.

3.6 Outcomes

The analysis uses four main classes of outcomes.

The first class is platform progression. These outcomes measure whether students started

the video, attempted Exercise 1, reached three correct answers in a row on Exercise 1, attempted Exercise 2, reached three correct answers in a row on Exercise 2, submitted the exit ticket, and completed the delayed assessment. These outcomes show how the treatments changed students' paths through the assignment.

The second class is practice behavior. For each exercise, we measure the number of questions completed, the number answered correctly, whether the student reached three correct answers in a row, and time spent. We also examine attempt-by-attempt outcomes within Exercise 1 among mastery students, including whether students reached later attempts and whether they answered correctly conditional on reaching them.

The third class is post-mistake recovery. For each mistake event, we measure whether the next attempt was correct, the number of attempts needed to return to a correct answer, whether the student eventually reached the three-correct-in-a-row threshold, and the amount of clock time needed to return to a correct answer. These outcomes are central for understanding the mechanism by which AI might affect learning.

The fourth class is delayed-test performance. The delayed assessment contained four questions: one question for each exercise in each of the two grade-specific topic bundles. For students assigned to a given topic, two questions corresponded to practiced material and two corresponded to unpracticed material. The main delayed-test outcomes are total practiced correct, total unpracticed correct, exercise-specific practiced correctness, and practiced-minus-unpracticed differences. Missing delayed-test items are coded as incorrect among students who took the delayed assessment.

3.7 Empirical specification

The main experimental estimates come from intent-to-treat regressions. For outcomes in the full experiment, the preferred specification is

$$Y_i = \alpha + \beta_1 AI_i + \beta_2 Mastery_i + \beta_3 (AI_i \times Mastery_i) + \varepsilon_i, \tag{1}$$

where AI_i indicates assignment to the NUMI AI tutor and $Mastery_i$ indicates assignment to mastery progression. The omitted group is CAL-only, non-mastery students. The coefficient β_1 estimates the effect of AI in the non-mastery condition, β_2 estimates the effect of mastery in the CAL-only condition, and β_3 estimates whether the effect of AI differs under mastery. The sum $\beta_1 + \beta_3$ gives the effect of AI within the mastery condition.

For the main AI mechanism analysis, we focus on mastery students and estimate

$$Y_i = \alpha + \delta AI_i + \varepsilon_i, \tag{2}$$

where δ compares AI to CAL-only support within the mastery workflow. This comparison is especially useful for Exercise 1 because all mastery students face the same progression rule and Exercise 1 is least affected by differential exposure to later material.

4 Results: Progression, Practice, and Delayed Learning

4.1 Sample composition and progression through the platform

Table 1 reports sample composition by grade, topic assignment, mastery assignment, and AI assignment. Randomization occurred at the individual level when students registered for the platform. Topic and AI assignment were approximately balanced across grades. Mastery assignment is also experimentally assigned, though the realized delayed-test sample differs somewhat across mastery status.

Table 3 shows how students progressed through the activity. Nearly all students reached and attempted Exercise 1, but progression to later parts of the platform was less complete. This pattern matters for interpretation. Exercise 1 is observed for almost everyone who begins the activity. Exercise 2, the exit ticket, and the delayed assessment are observed for smaller and potentially more selected groups. The treatments themselves also affect this progression, making later outcomes harder to interpret as pure learning outcomes.

Figure 1 depicts progression funnel across the main platform milestones. The figure shows, for each of the four AI-by-mastery cells, the share of students who attempted Exercise 1, reached three correct in a row on Exercise 1, attempted Exercise 2, reached three correct in a row on Exercise 2, and submitted the exit ticket. The main takeaway is that the treatments changed exposure. Students in the four groups did not simply experience the same assignment with different support; they reached different parts of the assignment at different rates.

4.2 Effects on practice behavior

Table 4 reports the effects of mastery, AI, and their interaction on practice behavior in Exercise 1 and Exercise 2. The omitted category is non-mastery students assigned to CAL-only support.

The first result is that mastery had large effects on Exercise 1 behavior. In the CAL-only condition, mastery increased the number of Exercise 1 questions completed by 4.62, increased the number answered correctly by 1.23, and increased the probability of reaching three correct answers in a row by 28.7 percentage points. Mastery also increased time spent on Exercise 1 by 6.69 minutes. These estimates show that the mastery rule substantially changed student behaviour in the platform. Students assigned to mastery practiced more and were much more likely to satisfy the platform's mastery threshold.

The second result is that AI changed the pace of practice. In the non-mastery condition, AI reduced the number of Exercise 1 questions completed by 1.04 and increased time spent by 1.64 minutes. The interaction between mastery and AI further reduced the number of Exercise 1 questions completed by 1.85, relative to what would be predicted from adding the two main effects. Thus, AI especially slowed progress inside the mastery workflow. The same broad pattern appears in Exercise 2: mastery increased practice and three-correct-in-a-row attainment, while AI reduced the number of questions completed and answered correctly, reflecting in part the fact that AI students progressed more slowly through the assignment.

These practice results highlight a central tradeoff. Mastery increased apparent completion. AI slowed students down. The combination of AI and mastery created a more intensive Exercise 1 experience, but also reduced the number of later questions students reached within the fixed class period. Therefore, Exercise 2 and exit-ticket outcomes are harder to interpret. They reflect not only what students learned, but also whether they reached the content and how much time remained once they did.

4.3 Mastery increased platform-defined success

Figure 2 highlights the most direct effect of the mastery rule. Mastery students were more likely to reach three correct answers in a row in both Exercise 1 and Exercise 2. This is expected, since the platform required mastery students to satisfy this threshold before progressing. However, the magnitude is important. In Exercise 1, the three-correct-in-a-row rate rose by almost 30 percentage points in the CAL-only group. In Exercise 2, mastery also increased the probability of reaching the threshold, despite the fact that fewer students reached later parts of the assignment.

The key question is whether this platform-defined success translated into delayed learning. If three correct answers in a row were a reliable measure of durable understanding, then the large increase in three-correct-in-a-row attainment should show up as improved performance one week later. The delayed-test results suggest a more cautious conclusion.

4.4 Delayed-test learning effects

Table 5 reports the main delayed-test outcomes from the full 2×2 design. The combined outcome is the practiced-minus-unpracticed delayed-test score. Exercise 1 and Exercise 2 report the corresponding practiced-minus-unpracticed differences separately by exercise position. The final two columns report total practiced and unpracticed scores.

The baseline practice effect is positive. Among non-mastery students assigned to CAL-only support, students answered more practiced than unpracticed delayed-test questions correctly. The combined practiced-minus-unpracticed difference is 0.108 questions, with 0.043 from Exercise 1 and 0.065 from Exercise 2. This confirms that the practice activity produced some retained learning relative to the unpracticed topic.

Mastery alone did not add to this delayed learning. The mastery coefficient is close to zero for the combined outcome, Exercise 1, Exercise 2, practiced score, and unpracticed score. This is the main mastery puzzle. The mastery rule substantially increased practice and three-correct-in-a-row attainment, but did not by itself produce detectable delayed-test gains.

The AI main effect in the non-mastery condition is also close to zero. However, the interaction between mastery and AI is mildly positive. For the combined practiced-minus-unpracticed score, the Mastery $\times$ AI estimate is 0.04; for Exercise 1 it is 0.037. These estimates are imprecise. The clearest positive estimate in the full table is for practiced delayed-test performance: the Mastery $\times$ AI coefficient is 0.085 on a two-question outcome, equivalent to 4.2 percentage points per question, though this estimate should be interpreted alongside the less precise practiced-minus-unpracticed contrast. This pattern is consistent with the hypothesis that AI is more useful when embedded in a structured workflow, but the full combined outcome is not precise enough to support a large or uniform learning effect.

Figure 3 summarizes the contrast between platform-defined success and delayed learning. The left panel shows that mastery increased the probability of reaching three correct answers in a row. The right panel shows that mastery alone did not increase delayed-test performance. This contrast is central for interpreting the study. Reaching a short-run streak during practice is not the same as retaining the material one week later.

This result provides suggestive evidence that a short-run threshold is an imperfect proxy for durable learning. The marginal students induced by the rule to reach three correct answers in a row may differ from students who would have reached the threshold anyway. For some students, reaching the threshold may reflect understanding. For others, it may reflect repeated exposure, short-run task success, or guessing through similar questions until a streak

is reached. The experiment cannot separately identify these channels. The randomized result is simpler: increasing platform-defined mastery did not, by itself, increase delayed learning.

The full experiment reveals two findings. First, mastery did what it was designed to do mechanically: it increased practice and three-correct-in-a-row attainment. Second, AI changed the pace of practice and created a time cost. These two facts make the later parts of the assignment more difficult to interpret. By Exercise 2, treatment effects combine learning, progression, and remaining time. Therefore, the next section focuses on the cleanest comparison for understanding the AI tutor's value-added: AI versus CAL-only support among mastery students on Exercise 1. Exercise 1 within the mastery workflow is the setting in which students share the same progression rule and exposure is least contaminated by downstream selection. It is also the setting most closely aligned with the mechanism: mastery creates a reason to persist after mistakes, while AI provides structured support at those moments.

5 AI Tutor Effects Within the Mastery Workflow

The full experiment shows that mastery and AI changed students' paths through the assignment. This section focuses on the cleaner comparison for understanding the AI tutor's value-added: students assigned to mastery, comparing AI to CAL-only support in Exercise 1. This is the setting in which all students faced the same progression rule and where exposure is least affected by differential progress to later material.

The analysis in this section has three parts. First, we show that AI changed the pace-quality tradeoff during Exercise 1. Students assigned to AI moved more slowly and reached fewer later questions, but made fewer observed mistakes. Second, we show that the strongest mechanism appears after mistakes: AI helped students recover with fewer additional attempts, though over more clock time. Third, we examine delayed-test performance, where the most encouraging learning evidence appears on practiced Exercise 1 material.

5.1 AI usage in the mastery condition

Table 6 summarizes AI usage among mastery students. Students assigned to AI in the mastery condition attempted 9.74 Exercise 1 questions on average. They used the "Help me get started" button 1.99 times, experienced 2.28 post-mistake walkthroughs, and clicked for 3.23 step explanations. About 42 percent typed at least one substantive math-related message to NUMI.

These usage rates matter for interpretation. AI access in NUMI was not merely a passive chat window that most students ignored. The interface deliberately brought the tutor into the practice workflow: students could request first-step help before attempting a question, and after mistakes the tutor intervened with a structured walkthrough or step-specific support. The estimates below therefore capture the effect of embedding AI into the practice environment, not simply making a generic chatbot available in the background.

5.2 AI changed the pace-quality tradeoff

Figure 4 shows the key tradeoff. The first panel plots the share of mastery students reaching three correct answers in a row over time. The second panel plots the number of questions attempted over time. The third panel plots correctness by attempt number among students who reached each attempt. The pattern is consistent across the panels: students assigned to AI progressed more slowly, but conditional on reaching a question, they were more likely to answer correctly.

A complication in interpreting attempt-level outcomes is that reaching later questions is itself an outcome affected by treatment. Conditional correctness among students who reach Question 5, for example, need not include the same types of students across AI and CAL-only groups. For this reason, Table 7 emphasizes reduced-form outcomes defined for all mastery students: whether the student reached each question and whether the student made a mistake on each question, coding students who did not reach the question as not making a mistake on that question. These estimates do not isolate correctness conditional on identical exposure. Instead, they describe the combined effect of AI on exposure and errors during the exercise.

Table 7 quantifies this pattern using attempt-by-attempt outcomes in Exercise 1. The first panel reports whether students reached each attempt. The second panel reports whether students made a mistake on each attempt. The two panels together provide suggestive evidence that AI students were less likely to reach later attempts but also less likely to make observed mistakes. By Question 5, AI reduced the probability of attempting the question by 10.4 percentage points and reduced the probability of making a mistake by 10.8 percentage points (comparing those who actually attempted the question).

5.3 AI helped students recover after mistakes

Mistakes are the moments when tutoring should matter most. A student who answers correctly may not need much help. A student who answers incorrectly must decide what

to do next. Try again quickly, look at a worked solution, ask for help, or give up. NUMI was designed to intervene at this point. After a first mistake, students assigned to AI were guided through the solution step by step. For subsequent mistakes, they alternated between solution review and additional walkthrough support.

Post-mistake outcomes raise a related selection issue because the sample is restricted to mistakes, and AI itself affects the probability and timing of mistakes. These estimates should therefore be interpreted as mechanism evidence rather than as the primary causal effect of AI on learning. They answer the question: when a mistake occurs in the mastery workflow, how does assignment to AI change what happens next? The randomized delayed-test outcomes remain the main evidence on learning.

Table 8 reports post-mistake recovery outcomes among mastery students. Each observation is a mistake event. The outcomes measure whether the next attempt was correct, how many additional attempts were needed to reach the next correct answer, whether the student eventually reached three correct answers in a row after the mistake, and how much clock time elapsed before the next correct answer.

AI improved recovery on the next attempt. Among mastery students who made a mistake, AI increased the probability that the next attempt was correct by 8.5 percentage points. AI also reduced the number of attempts needed to reach the next correct answer by 0.96 attempts. These are large effects relative to the immediate post-error setting. At the same time, AI increased the amount of clock time needed to reach the next correct answer by 2.88 minutes.

The post-mistake results provide the strongest evidence on mechanism, subject to the conditioning described above. They show that, for observed mistakes in the mastery workflow, AI changed what happened at the moment when students were most likely to need help. The tutor did not solely increase practice volume. It altered the process of recovering from errors, making the next answer more likely to be correct and reducing the number of additional attempts needed to get back on track.

5.4 Is the post-mistake effect just the first-step help button?

One concern is that the post-mistake results may partly reflect students receiving help before making their initial attempt. If a student used the "Help me get started" button before answering, the student may have been better positioned to recover after a subsequent error. This would still be part of the AI treatment, but it would be useful to know whether the post-mistake effect is driven entirely by pre-answer first-step help or by the tutor's post-error

support.

We address this concern in the appendix with a help-channel decomposition that subtracts the estimated direct and indirect contributions of first-step help from the observed post-error AI gaps (Appendix Table A10). The direct channel is a student's use of *Help me get started* on the next question itself; the indirect channel is selection into the mistake-conditioned sample, since earlier hint use may have prevented the triggering mistake. These adjustments are more model-dependent than the experimental intent-to-treat estimates, so we treat them as robustness checks rather than primary evidence. For the transition from a wrong answer on Question 1 to a correct answer on Question 2, most of the small observed advantage (4.3 percentage points) is consistent with the first-step hint channel: the adjusted gap is not distinguishable from zero. For the transition from a wrong answer on Question 2 to a correct answer on Question 3 among mastery students, a post-error advantage survives even the most conservative correction: the observed AI gap of 9.4 percentage points falls to 5.9 percentage points but remains significant at the 5 percent level, even under the assumption that every student kept out of the mistake-conditioned sample by *Help me get started* on Question 2 would have answered Question 3 incorrectly. We read this as evidence that the recovery pattern is not entirely reducible to the hint button, while acknowledging that the hint channel accounts for much of the earliest post-error gap.

The main interpretation does not hinge on this decomposition. The experimental AI treatment bundled first-step and post-mistake support, and both are legitimately part of the effect. The decomposition serves only to locate how much of the observed recovery pattern is attributable to the post-error walkthrough specifically, and it suggests that at least at the later transition, the walkthrough adds something beyond the initial hint.

5.5 Delayed learning among mastery students

Table 9 reports delayed-test effects of AI among mastery students. The cleanest outcome is practiced Exercise 1 performance, because Exercise 1 was the content most consistently reached by mastery students and the focus of the post-mistake analysis.

Among mastery students, CAL-only students answered 37.0 percent of the practiced Exercise 1 delayed-test question correctly. Students assigned to AI answered 40.2 percent correctly, a difference of 3.1 percentage points. The corresponding effect on the unpracticed Exercise 1 question is close to zero. Thus, the Exercise 1 pattern is consistent with a learning effect on practiced material rather than a general difference in test-taking or topic difficulty. The estimate is imprecise, with a p-value of 0.065 in the current table, so it should be

interpreted as suggestive.

Figure 5 shows the practiced Exercise 1 result visually, alongside the corresponding unpracticed Exercise 1 outcome. The effect is not transformational. Most students still answered the delayed question incorrectly. But the magnitude is meaningful relative to the amount of practice. In the full delayed-test table, the practiced-minus-unpracticed gap for CAL-only non-mastery students is about 4.3 percentage points for Exercise 1. The AI effect among mastery students on practiced Exercise 1 is about 3.1 percentage points. Despite the estimates being imprecise, the AI value-added is similar in size to the gain from practice itself for this exercise.

The Exercise 2 delayed-test results are harder to interpret. AI and mastery affected whether students reached Exercise 2 and how much time they had once they got there. A lower or noisier Exercise 2 effect could reflect less exposure, different selection into reaching Exercise 2, or a true difference in learning. For this reason, Exercise 2 results are reported in the full tables and appendix, but are not the cleanest evidence on AI value-added.

Taken together, the mastery-sample results support a coherent interpretation. AI slowed students down and reduced exposure to later questions. But during the shared Exercise 1 workflow, it improved observed accuracy and helped students recover after mistakes. The delayed-test evidence is consistent with modest learning gains on the practiced material most closely tied to that mechanism. The estimates are not large enough or precise enough to claim that a short AI-supported practice session substantially transforms learning. They do show that a guard-railed LLM tutor can add value over a CAL-only platform when embedded in a structure that makes post-mistake support salient.

5.6 Robustness

Several additional analyses reinforce this interpretation. First, the AI effects appear stronger on the harder questions. This is useful because the theory is that AI support should matter most when students are uncertain, make mistakes, or need help identifying the next step. If the positive AI estimates are concentrated on harder or more open-ended items, that would strengthen the interpretation that NUMI helped students reason through material rather than merely benefit from easier question formats.

Second, the results are stronger in periods when the platform experienced fewer lags. During the first week of the experiment, website performance was slower in some earlier sessions from heavy usage. The issue was corrected later in the week. The larger effect estimates during sessions with fewer NUMI time delays matters because the AI treatment

required more back-and-forth interaction with the platform than the CAL-only condition. Slowness should therefore attenuate the AI treatment more than the CAL-only treatment. Appendix Table A2 (Panel C) reports robustness checks splitting the sample by platform performance. The post-mistake results are at least as strong on the lower-lag days: for example, the AI coefficient on next-attempt correctness is larger on Thursday and Friday than on Tuesday and Wednesday, while the reduction in attempts to the next correct answer remains negative in both samples. Delayed-test estimates also become more favorable in samples with lower measured wait times. These patterns are consistent with implementation frictions attenuating, rather than generating, the main AI effects.

Third, delayed-test effects are more positive among students who cleared the first mastery gate, although these estimates are descriptive because clearing the gate is endogenous. The purpose of this exercise is to study whether the delayed-test pattern is stronger among students who actually experienced the intended mastery workflow. In the conditional tables, the AI effect among mastery students is larger for practiced Exercise 1 among students who completed Exercise 1 than in the full mastery sample. This pattern is consistent with the interpretation that AI value-added is most visible when students both receive the AI-supported practice and make enough progress through the mastery sequence for the support to matter.

Taken together, these robustness checks point in the same direction as the main mechanism results. The AI effects are not strongest where one would expect pure answer-getting or superficial clicking to matter most. They are stronger where the platform worked more smoothly, where questions were more challenging, and where students experienced the intended mastery workflow. Because the latter two analyses condition on realized post-randomization behavior or item characteristics, they should not replace the randomized estimates. But they help explain why the main delayed-learning effects are modest in the full sample and why the strongest evidence appears in the mastery Exercise 1 setting.

6 Interpretation and Limitations

The results suggest a useful way to think about AI tutoring in classroom practice. AI access by itself is not the only object of interest. What matters is the combination of the tutor, the platform, and the incentives students face while practicing. NUMI was designed around this idea. The AI tutor was not a separate chatbot that students could use if they wanted. It was embedded in the practice environment, appeared at moments when students needed help, and was paired with a mastery rule that made mistakes more consequential.

The main result is that students assigned to AI progressed more slowly and reached fewer later questions. This is an important limitation, but it is also informative. If an AI tutor is doing more than giving answers, it may take time. The question is whether this additional time is productive. The post-mistake results suggest that at least some of it was. Students assigned to AI were more likely to answer correctly after mistakes and needed fewer additional attempts to return to a correct answer, even though more clock time elapsed. This is the sense in which AI created a productive slowdown.

The mastery results are useful partly because they complicate a simple story. Requiring students to reach three correct answers in a row substantially increased platform-defined mastery. It also increased practice and the number of correct answers during practice. Yet it did not, by itself, increase delayed-test performance. A short-run streak therefore should not be treated as equivalent to durable understanding.

There are several possible explanations. Some students may have reached the threshold because they understood the material. Others may have reached it through repeated exposure to similar problems, lucky streaks, or guessing until a sequence worked. Some may have understood the material during practice but not retained it one week later. Others may not have taken the delayed assessment as seriously as the practice activity, although the assessment was presented as formal and students were instructed to do their best. The experiment does not separately identify these channels.

Nevertheless, the randomized result is clear. Increasing the probability that students reached a three-correct-in-a-row threshold did not, by itself, improve delayed learning. This does not mean that mastery learning is ineffective. It means that the particular platform implementation of mastery used here was not sufficient on its own. A progression rule can create persistence, but persistence is not automatically productive. What happens after a student makes a mistake appears to matter.

This helps explain why AI and mastery may be complements. Mastery creates a reason to keep working after a mistake. AI changes what happens during that continued work. If a student must continue practicing but has only a worked solution to consult, the additional attempts may or may not produce understanding. If the same student receives an interactive explanation, a first-step reasoning prompt, or a post-error walkthrough, the additional time may become more useful. The evidence in this paper is consistent with that interpretation.

The most important implementation tradeoff is time. AI support slowed students down. In a fixed class period, time spent with the tutor can reduce the number of questions attempted and the probability of reaching later content. This is especially relevant for Exercise 2. Students assigned to AI within mastery spent more time on Exercise 1 and were less likely

to experience later parts of the assignment in the same way as CAL-only students. As a result, Exercise 2 outcomes combine several effects: learning from AI support, reduced exposure, differences in remaining time, and selection into reaching the exercise.

This tradeoff is not unique to our study. Any tutoring intervention uses time. A human tutor who asks a student to explain a misconception also slows the student down relative to simply revealing an answer. The policy question is not whether the intervention saves time, but whether it uses time well. In this experiment, the answer appears mixed but encouraging. The additional time did not produce large delayed-test gains in the full sample. But it improved post-mistake recovery and produced the clearest delayed-learning signal on the practiced Exercise 1 material most closely tied to the AI-supported workflow.

Future versions of AI tutoring platforms will need to manage this tradeoff more carefully. One possibility is to use AI support more selectively, intervening most intensively when the student is likely to be stuck or when the item is especially diagnostic. Another is to design shorter but more targeted post-error interactions. A third is to adapt the mastery rule so that students receive enough support to learn from mistakes without spending so much time on early exercises that they fail to reach later material.

Finally, continued improvements in AI capability may shrink the tradeoff itself. A more capable tutor could deliver the same instruction in less time, leaving students more room to reach later material.

6.1 Limitations

A few limitations should be kept in mind in interpreting this study's results.

First, the intervention duration was short. The experiment studied one NUMI assignment and one delayed assessment approximately one week later. The study does not estimate effects on course grades, state assessment scores, or longer-run math achievement. A natural next step is to study repeated use over multiple weeks or units.

Second, the delayed assessment was brief. Four questions allow a clean practiced-versus-unpracticed comparison, but they also produce noisy outcome measures. This makes it harder to detect modest learning effects and harder to study heterogeneity by topic, question type, or prior achievement.

Third, Exercise 2 outcomes are difficult to interpret because treatment affected exposure. This possibility was pre-registered, but it still limits what can be concluded from later-exercise outcomes.

Fourth, the mastery rule was only one possible way to structure AI use. Requiring three correct answers in a row is simple and easy to implement, but it may be too weak, too gameable, or too disconnected from conceptual understanding. A stricter rule, such as five correct answers in a row, might reduce the chance of reaching the threshold through guessing, but it would also increase time costs and further reduce exposure to later material. More promising alternatives may involve varied item types, brief explanation requirements, delayed retrieval checks, or AI-triggered misconception repair.

Fifth, NUMI was a specific platform, not generic AI access. This is a strength for internal validity, because the AI was guard-railed and embedded in a defined learning environment. But it also means the results should not be generalized to all uses of LLMs in education. The effect of AI depends on the interface, prompts, timing of support, content quality, and student incentives.

7 Conclusion

Large language models have generated enormous interest as potential tutors. The key question is not whether they can produce plausible explanations. They can. The harder question is whether students use those explanations in ways that improve learning.

This paper studies that question in a randomized field experiment with more than 6,000 middle-school students using NUMI, a purpose-built computer-assisted learning platform. Students were randomized to AI versus CAL-only support, mastery versus non-mastery progression, and one of two math topics. The results show that AI changed the production of practice. It slowed students down and reduced exposure to later questions, but improved observed accuracy and post-mistake recovery within the mastery workflow. Mastery alone increased three-correct-in-a-row attainment but did not improve delayed learning. The most encouraging delayed-test evidence appears when AI is embedded in the mastery workflow, with modest gains concentrated on practiced Exercise 1 material.

The results are not transformational, but they carry important policy implications. They provide evidence that an LLM-based tutor can add value over an otherwise similar CAL platform in real classroom math practice. They also show that the surrounding structure matters. AI support appears most useful when students have a reason to persist after mistakes and when the tutor helps make that persistence productive.

The results point to a design agenda. AI tutors appear most useful when they are not treated as optional help desks but embedded so that students have a reason to slow down

and use help at the right moment. Mastery rules are one way to create such reasons, but not the only one. Requiring students to explain an error, eliciting a confidence rating before feedback, or inserting a delayed retrieval check before declaring mastery may work as well or better. The underlying challenge is to create productive friction: too little, and students rush, guess, or use AI to obtain answers; too much, and they exhaust the class period on a few questions or disengage—and NUMI's results suggest that a well-designed tutor can make some of this friction productive by helping students recover after mistakes, even though the time costs are real. Future systems that are more conversational, multimodal, and aware of student work may ease this tradeoff by intervening when help is most valuable and staying out of the way when it is not. These remain design hypotheses. NUMI's contribution is to showcase that they can be tested at scale, moving beyond demonstrations of what AI can say toward evidence on what it helps students learn.

The broader finding from this study is that AI tutoring should not be evaluated as a stand-alone technology. It should be evaluated as part of a learning environment. The promise of AI in education will depend less on whether models can answer questions and more on whether platforms, teachers, and students use them to create environments that are conducive to learning.

References

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden, "Remedying Education: Evidence from Two Randomized Experiments in India," *Quarterly Journal of Economics*, 2007, *122* (3), 1235–1264.

Barrow, Lisa, Lisa Markman, and Cecilia Elena Rouse, "Technology's Edge: The Educational Benefits of Computer-Aided Instruction," *American Economic Journal: Economic Policy*, 2009, *1* (1), 52–74.

Bastani, Hamsa, Osbert Bastani, Alp Sungu, Haosen Ge, Ozge Kabakci, and Rei Mariman, "Generative AI Without Guardrails Can Harm Learning: Evidence from High School Mathematics," *Proceedings of the National Academy of Sciences*, 2025, *122* (26), e2422633122.

Bjork, Elizabeth L. and Robert A. Bjork, "Making Things Hard on Yourself, But in a Good Way: Creating Desirable Difficulties to Enhance Learning," *Psychology and the Real World*, 2011, pp. 56–64.

Bloom, Benjamin S., "Learning for Mastery," *Evaluation Comment*, 1968, *1* (2), 1–12.

Bulman, George and Robert W. Fairlie, "Technology and Education: Computers, Software, and the Internet," *Handbook of the Economics of Education*, 2016, *5*, 239–280.

Chi, Michelene T. H., "Self-Explaining Expository Texts: The Dual Processes of Generating Inferences and Repairing Mental Models," *Advances in Instructional Psychology*, 2000, *5*, 161–238.

Dietrichson, Jens, Martin Bog, Trine Filges, and Anne-Marie Klint Jorgensen, "Academic Interventions for Elementary and Middle School Students with Low Socioeconomic Status: A Systematic Review and Meta-Analysis," *Review of Educational Research*, 2017, *87* (2), 243–282.

Dunlosky, John, Katherine A. Rawson, Elizabeth J. Marsh, Mitchell J. Nathan, and Daniel T. Willingham, "Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology," *Psychological Science in the Public Interest*, 2013, *14* (1), 4–58.

Escueta, Maya, Vincent Quan, Andre Joshua Nickow, and Philip Oreopoulos, "Education Technology: An Evidence-Based Review," *Journal of Economic Literature*, 2020, *58* (4), 897–996.

Falkner, Nick, Juho Leinonen, Miranda C. Parker, Andrew Petersen, and Claudia Szabo, "A Game of Shadows: Effective Mastery Learning in the Age of Ubiquitous AI (Dagstuhl Seminar 24272)," *Dagstuhl Reports*, 2024, *14* (6), 245–262.

Fesler, Lily, JP Martinez Claeys, Chris Agnew, and Susanna Loeb, "The Evidence Base on AI in K-12: A 2026 Review," Technical Report, AI Hub for Education, SCALE Initiative, Stanford University 2026.

Ganimian, Alejandro J., Emiliana Vegas, and Frederick M. Hess, "Realizing the Promise: How Can Education Technology Improve Learning for All?," *Brookings Institution Report*, 2023.

Garner, James, Paul Denny, and Andrew Luxton-Reilly, "Mastery learning in computer science education," in "Proceedings of the Twenty-First Australasian Computing Education Conference" 2019, pp. 37–46.

Guskey, Thomas R, "Lessons of mastery learning," *Educational leadership*, 2010, *68* (2), 52–57.

Henkel, Owen, Hannah Horne-Robinson, Nessie Kozhakhmetova, and Amanda Lee, "Effective and Scalable Math Support: Experimental Evidence on the Impact of an AI-Math Tutor in Ghana," *Working paper*, 2024.

Kapur, Manu, "Productive Failure," *Cognition and Instruction*, 2008, *26* (3), 379–424.

Kestin, Greg, Kelly Miller, Anna Klales, Timothy Milbourne, and Gregorio Ponti, "AI Tutoring Outperforms In-Class Active Learning: An RCT Introducing a Novel Research-Based Design in an Authentic Educational Setting," *Scientific Reports*, 2025, *15*, 17458.

Koedinger, Kenneth R, Albert T Corbett, and Charles Perfetti, "The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning," *Cognitive science*, 2012, *36* (5), 757–798.

Kulik, Chen-Lin C., James A. Kulik, and Robert L. Bangert-Drowns, "Effectiveness of Mastery Learning Programs: A Meta-Analysis," *Review of Educational Research*, 1990, *60* (2), 265–299.

Kumar, Harsh, Ilya Musabirov, Joseph Jay Williams, and Michael Liut, "QuickTA: exploring the design space of using large language models to provide support to students," in "Learning Analytics and Knowledge Conference 2023 (LAK'23)" 2023.

_ , _ , Mohi Reza, Jiakai Shi, Xinyuan Wang, Joseph Jay Williams, Anastasia Kuzminykh, and Michael Liut, "Guiding Students in Using LLMs in Supported Learning Environments: Effects on Interaction Dynamics, Learner Performance, Confidence, and Trust," *Proc. ACM Hum.-Comput. Interact.*, November 2024, *8* (CSCW2).

Lehmann, Matthias, Philipp B. Cornelius, and Fabian J. Sting, "AI Meets the Classroom: When Does ChatGPT Harm Learning?," *arXiv preprint arXiv:2409.09047*, 2024.

Leinonen, Juho, Arto Hellas, Sami Sarsa, Brent Reeves, Paul Denny, James Prather, and Brett A Becker, "Using large language models to enhance programming error messages," in "Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1" 2023, pp. 563–569.

Liffiton, Mark, Brad E Sheese, Jaromir Savelka, and Paul Denny, "Codehelp: Using large language models with guardrails for scalable support in programming classes," in "Proceedings of the 23rd Koli calling international conference on computing education research" 2023, pp. 1–11.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J. Ganimian, "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India," *American Economic Review*, 2019, *109* (4), 1426–1460.

Nickow, Andre Joshua, Philip Oreopoulos, and Vincent Quan, "The Impressive Effects of Tutoring on PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence," EdWorkingPaper 20-267, Annenberg Institute at Brown University 2020.

Nie, Allen, Yash Chandak, Miroslav Suzara, Ali Malik, Juliette Woodrow, Matt Peng, Mehran Sahami, Emma Brunskill, and Chris Piech, "The GPT Surprise: Offering Large Language Model Chat in a Massive Coding Class Reduced Engagement but Increased Adopters' Exam Performances," *arXiv preprint arXiv:2407.09975*, 2024.

Pardos, Zachary A. and Shreya Bhandari, "ChatGPT-Generated Help Produces Learning Gains Equivalent to Human Tutor-Authored Help on Mathematics Skills," *PLOS ONE*, 2024, *19* (5), e0304013.

Renkl, Alexander, "Learning from Worked-Out Examples: Instructional Explanations Supplement Self-Explanations," *Learning and Instruction*, 2002, *12* (5), 529–556.

Ritter, Gary W., Joshua H. Barnett, George S. Denny, and Ginger R. Albin, "The Effectiveness of Volunteer Tutoring Programs for Elementary and Middle School Students: A Meta-Analysis," *Review of Educational Research*, 2009, *79* (1), 3–38.

Roediger, Henry L. and Jeffrey D. Karpicke, "Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention," *Psychological Science*, 2006, *17* (3), 249–255.

Sarsa, Sami, Paul Denny, Arto Hellas, and Juho Leinonen, "Automatic generation of programming exercises and code explanations using large language models," in "Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1" 2022, pp. 27–43.

Savelka, Jaromir, Paul Denny, Mark Liffiton, and Brad Sheese, "Efficient classification of student help requests in programming courses using large language models," *arXiv preprint arXiv:2310.20105*, 2023.

Vanzo, Alessandro, Sankalan Pal Chowdhury, and Mrinmaya Sachan, "GPT-4 as a Homework Tutor Can Improve Student Engagement and Learning Outcomes," *arXiv preprint arXiv:2409.15981*, 2024.

Vygotsky, Lev S., *Mind in Society: The Development of Higher Psychological Processes*, Harvard University Press, 1978.

Wang, Rose E., Ana T. Ribeiro, Carly D. Robinson, Susanna Loeb, and Dora Demszky, "Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise," *arXiv preprint arXiv:2410.03017*, 2025.

Wood, David, Jerome S. Bruner, and Gail Ross, "The Role of Tutoring in Problem Solving," *Journal of Child Psychology and Psychiatry*, 1976, *17* (2), 89–100.

Table 1: Sample Composition by Grade, Topic, Mastery Status, and AI Status

	Grade 6	Grade 7	Grade 8	Total
Total Number	2527	2463	2007	6997
Fraction Topic A	0.512	0.492	0.488	0.498
Fraction Topic B	0.488	0.508	0.512	0.502
Difference	0.024	-0.015	-0.023	-0.003
Fraction Mastery	0.520	0.509	0.499	0.510
Fraction Non-Mastery	0.480	0.491	0.501	0.490
Difference	0.041**	0.017	-0.001	0.020*
Fraction AI	0.500	0.490	0.476	0.490
Fraction CAL-only	0.500	0.510	0.524	0.510
Difference	-0.000	-0.020	-0.047**	-0.021*

Notes: This table describes the composition of the week-1 analysis sample by grade. The sample is restricted to students in grades 6–8 with non-missing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity. Columns report values separately for Grade 6, Grade 7, Grade 8, and the full sample. For each pair, the difference row reports the first-listed fraction minus the second-listed fraction. Statistical significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: Balance of Baseline Characteristics in the Delayed-Test Sample

Variable	AI + Mastery	AI + Non-mastery	CAL-only + Mastery	CAL-only + Non-mastery	F-test p-value	Nonmissing N
Female	0.487	0.493	0.473	0.495	0.647	5,752
Black	0.262	0.259	0.279	0.286	0.306	5,752
Hispanic	0.236	0.242	0.249	0.240	0.889	5,752
Free lunch	0.472	0.431	0.471	0.474	0.064	5,752
Economically disadvantaged	0.303	0.275	0.287	0.285	0.415	5,752
English learner	0.147	0.153	0.155	0.156	0.909	5,752
Student with disability	0.181	0.169	0.180	0.171	0.766	5,752
Age	13.086	13.124	13.146	13.155	0.161	5,752
ELA NWEA MAP RIT (z-score)	0.054	0.047	0.022	0.001	0.519	4,547
Math NWEA MAP RIT (z-score)	0.038	0.043	0.011	0.031	0.850	4,558

Notes: This table reports baseline balance across the four randomized treatment groups in the delayed-test analysis sample. The sample is restricted to students in grades 6–8 with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity. Free lunch is an indicator for FARMS coded "F." ELA and Math NWEA MAP RIT scores are winter 2025–26 RIT scores standardized using the full nonmissing 2025–26 assessment population. The F-test p-value is from a test of equality of means across the four treatment groups.

Table 3: Week-1 Platform Engagement and Subsequent Participation by Grade, Topic, and Mastery Status

	Week-1 Login	Started Video 1	Watched 75% of Video 1	Attempted Exercise 1	3 Correct on Ex. 1	Started Video 2	Watched 75% of Video 2	Attempted Exercise 2	3 Correct on Ex. 2	Submitted Ticket	Took Test Week Later
Panel A. Total Sample											
Grade 6 Topic A	1294	0.99	0.92	0.98	0.26	0.57	0.46	0.50	0.16	0.34	0.89
Grade 6 Topic B	1233	0.99	0.97	0.99	0.51	0.81	0.76	0.74	0.25	0.52	0.91
Grade 7 Topic A	1213	1.00	0.97	0.98	0.52	0.83	0.74	0.76	0.20	0.46	0.93
Grade 7 Topic B	1250	0.99	0.97	0.97	0.35	0.66	0.53	0.62	0.30	0.51	0.92
Grade 8 Topic A	980	0.99	0.95	0.96	0.41	0.69	0.51	0.59	0.23	0.44	0.89
Grade 8 Topic B	1027	0.99	0.95	0.98	0.58	0.84	0.67	0.79	0.40	0.66	0.88
Total	6997	0.99	0.96	0.98	0.43	0.73	0.61	0.66	0.25	0.48	0.90
Panel B. Mastery Sample											
Grade 6 Topic A	678	1.00	0.95	0.98	0.37	0.37	0.28	0.32	0.20	0.20	0.90
Grade 6 Topic B	637	0.99	0.98	0.98	0.68	0.70	0.67	0.64	0.32	0.39	0.89
Grade 7 Topic A	616	1.00	0.98	0.99	0.69	0.78	0.71	0.69	0.28	0.33	0.94
Grade 7 Topic B	637	1.00	0.99	0.97	0.47	0.51	0.42	0.48	0.36	0.40	0.91
Grade 8 Topic A	481	1.00	0.96	0.96	0.55	0.58	0.46	0.50	0.32	0.35	0.88
Grade 8 Topic B	521	1.00	0.96	0.98	0.74	0.78	0.62	0.74	0.52	0.58	0.88
Total	3570	1.00	0.97	0.98	0.58	0.61	0.52	0.56	0.33	0.37	0.90
Panel C. Non-Mastery Sample											
Grade 6 Topic A	616	0.98	0.90	0.97	0.15	0.78	0.65	0.70	0.12	0.50	0.89
Grade 6 Topic B	596	0.99	0.97	0.99	0.33	0.92	0.85	0.85	0.18	0.65	0.93
Grade 7 Topic A	597	0.99	0.96	0.98	0.34	0.88	0.78	0.82	0.11	0.59	0.92
Grade 7 Topic B	613	0.99	0.96	0.96	0.23	0.82	0.65	0.76	0.23	0.62	0.93
Grade 8 Topic A	499	0.99	0.94	0.96	0.28	0.81	0.57	0.69	0.14	0.54	0.89
Grade 8 Topic B	506	0.98	0.93	0.98	0.41	0.90	0.71	0.85	0.28	0.74	0.89
Total	3427	0.99	0.94	0.97	0.29	0.85	0.70	0.78	0.18	0.60	0.91

Notes: This table reports student progression through the platform during the first week and participation in the delayed assessment administered approximately one week later. The sample is restricted to students in grades 6–8 with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity. Panel A reports the total week-1 sample, Panel B reports the mastery subsample, and Panel C reports the non-mastery subsample. Each row shows the number of students who logged into the platform in week 1 and, for all subsequent columns, the fraction of students in the row group who reached the indicated milestone. "Watched 75% of Video 1" and "Watched 75% of Video 2" indicate whether the student watched at least 75 percent of the corresponding instructional video. "Attempted Exercise 1" and "Attempted Exercise 2" indicate whether the student attempted at least one question in the corresponding exercise. "3 Correct on Ex. 1" and "3 Correct on Ex. 2" indicate whether the student answered three questions in a row correctly in the corresponding exercise. "Submitted Ticket" indicates submission of the week-1 exit ticket. "Took Test Week Later" indicates that the student took the delayed assessment the following week.

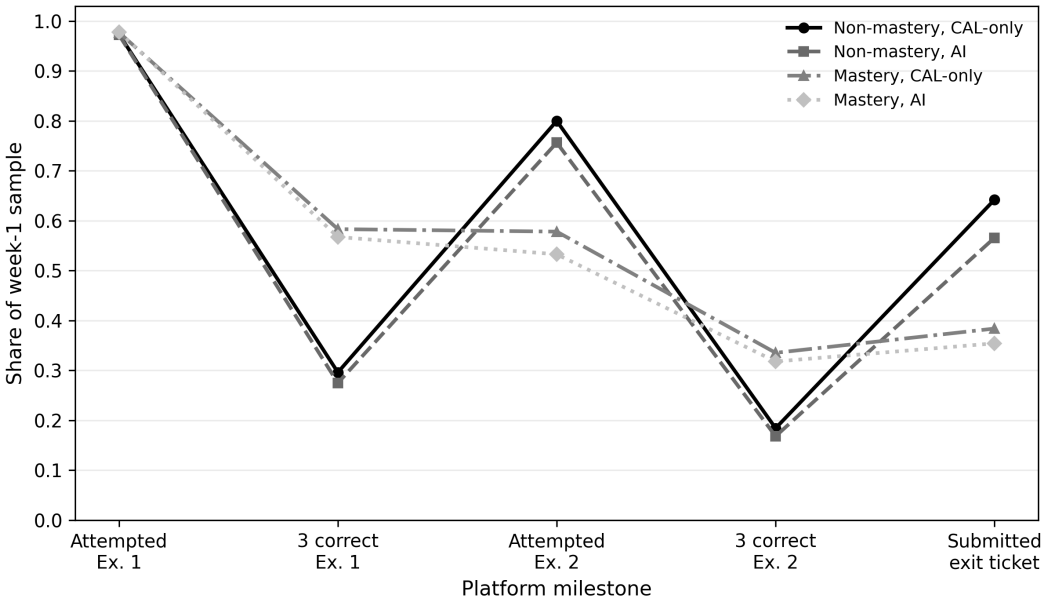


Figure 1: Progression Through the Week-1 NUMI Activity by Treatment Arm

Notes: This figure reports the fraction of students in each treatment arm reaching selected milestones in the week-1 NUMI activity. The sample is restricted to students in grades 6–8 with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity. The milestones are sequential platform outcomes: attempting Exercise 1, answering three questions in a row correctly on Exercise 1, attempting Exercise 2, answering three questions in a row correctly on Exercise 2, and submitting the exit ticket. The figure excludes week-2 delayed-test participation because the delayed assessment was administered separately and is not a sequential week-1 platform milestone.

Table 4: Exercise 1 and Exercise 2 Outcomes by Mastery Status and AI Assignment

	Questions Completed	Number Correct	3 Correct in a Row	Exercise Time
Panel A. Exercise 1				
Mastery	4.615***	1.232***	0.287***	6.692***
	(0.255)	(0.095)	(0.016)	(0.444)
AI	-1.042***	-0.161*	-0.021	1.643***
	(0.144)	(0.090)	(0.015)	(0.391)
Mastery × AI	-1.850***	-0.149	0.006	-0.453
	(0.307)	(0.125)	(0.023)	(0.639)
CAL-only/non-mastery mean	5.223	2.240	0.296	8.989
Observations	6,997	6,997	6,997	6,997
Panel B. Exercise 2				
Mastery	1.067***	0.466***	0.151***	1.446***
	(0.190)	(0.073)	(0.014)	(0.327)
AI	-0.983***	-0.252***	-0.016	-0.143
	(0.108)	(0.062)	(0.013)	(0.270)
Mastery × AI	-0.750***	-0.095	-0.001	-0.395
	(0.225)	(0.097)	(0.020)	(0.461)
CAL-only/non-mastery mean	3.462	1.422	0.184	4.922
Observations	6,997	6,997	6,997	6,997

Notes: Each column reports a separate intent-to-treat regression of the indicated exercise-specific outcome on mastery assignment, AI assignment, and their interaction. The omitted category is students assigned to non-mastery progression and CAL-only support. Panel A reports outcomes for Exercise 1 and Panel B reports outcomes for Exercise 2. "Questions Completed" is the number of exercise questions with at least one non-skip submitted attempt. "Number Correct" is the number of exercise questions answered correctly. "3 Correct in a Row" is an indicator equal to one if the student answered three questions consecutively correctly in the corresponding exercise. "Exercise Time" is measured in minutes as the elapsed time between the first and last non-skip submitted attempt within the exercise, capped at 60 minutes; students with no non-skip submitted attempt in the exercise are coded as zero. This timing measure does not include video time, time before the first submitted attempt, time after the last submitted attempt, exit-ticket time, or other platform time outside the exercise-attempt window. The sample is restricted to students in grades 6–8 with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity. Missing exercise outcomes are coded as zero. Robust standard errors are reported in parentheses. Statistical significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

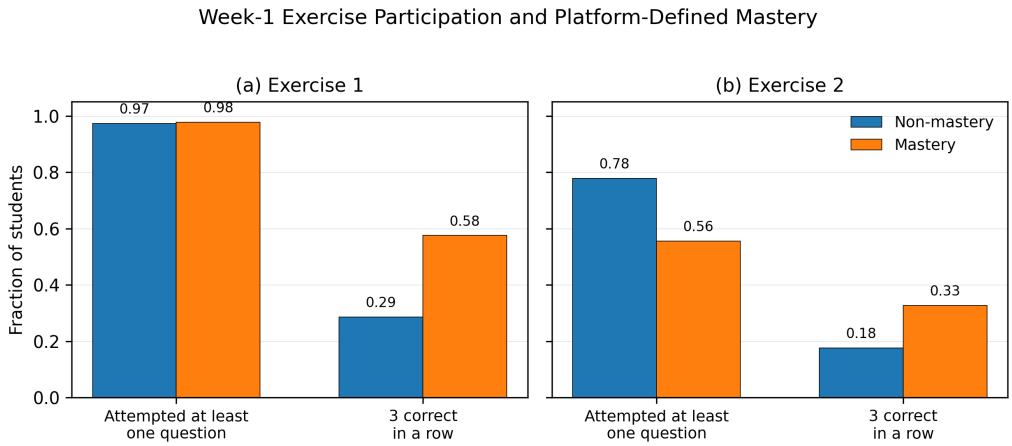


Figure 2: Exercise Participation and Platform-Defined Mastery by Mastery Assignment

Notes: This figure reports week-1 platform progression by mastery assignment. Panel (a) reports Exercise 1 and Panel (b) reports Exercise 2. Within each panel, the first category shows the fraction of students who attempted at least one question in the exercise, and the second category shows the fraction who attained three correct answers in a row within the exercise. Bars are labeled with the corresponding sample fractions. The sample is restricted to students in grades 6–8 with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity. Missing exercise progression outcomes are coded as zero.

Table 5: Main Delayed-Test Learning Effects

	Combined	Exercise 1	Exercise 2	Practiced	Unpracticed
Mastery	-0.010	-0.012	0.001	-0.040	-0.029
	(0.034)	(0.023)	(0.021)	(0.028)	(0.026)
AI	-0.014	-0.007	-0.007	-0.041	-0.027
	(0.034)	(0.023)	(0.022)	(0.028)	(0.027)
Mastery × AI	0.040	0.037	0.002	0.085**	0.046
	(0.048)	(0.032)	(0.030)	(0.039)	(0.038)
CAL-only/non-mastery mean	0.108	0.043	0.065	0.780	0.672
Observations	6,327	6,327	6,327	6,327	6,327

Notes: Each column reports a separate intent-to-treat regression of the indicated delayed-test outcome on mastery assignment, AI assignment, and their interaction. The omitted category is students assigned to non-mastery progression and CAL-only support. The combined outcome is the practiced-minus-unpracticed delayed-test score. Exercise 1 and Exercise 2 report the practiced-minus-unpracticed difference separately by exercise position. Practiced and unpracticed outcomes are total correct out of two questions. The sample is restricted to students in grades 6–8 with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity and taking the week-2 delayed assessment. Missing delayed-test item responses are coded as incorrect among delayed-test takers. Robust standard errors are reported in parentheses. Statistical significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

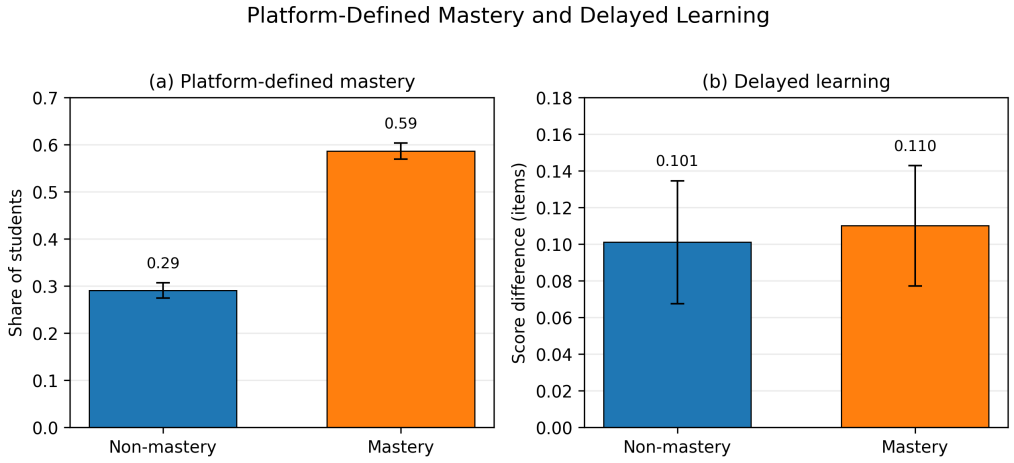


Figure 3: Platform-Defined Mastery and Delayed Learning

Notes: This figure compares students assigned to mastery progression with students assigned to non-mastery progression. Both panels use the final delayed-test analysis sample: students in grades 6–8 with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity and taking the week-2 delayed assessment. Panel (a) reports the fraction of students who answered three questions in a row correctly on Exercise 1. Panel (b) reports the combined practiced-minus-unpracticed delayed-test score, defined as total correct on the two practiced delayed-test items minus total correct on the two unpracticed delayed-test items. Missing delayed-test item responses are coded as incorrect among delayed-test takers. Bars report group means pooling across AI assignment. Error bars show 95 percent confidence intervals for the group mean. Values above bars report the corresponding group means.

Table 6: AI Usage Among Mastery Students

Outcome	Mean among AI mastery students
Total questions attempted	9.74
"Help me get started" uses	1.99
Post-mistake walkthroughs	2.28
Step-explanation clicks	3.23
Typed substantive math-related message	0.42

Notes: This table reports average AI usage among students assigned to both AI support and mastery progression. The sample is restricted to students in grades 6–8 with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity, assigned to mastery progression, and assigned to AI support. "Total questions attempted" counts the total number of Exercise 1 and Exercise 2 questions with at least one submitted attempt. "Help me get started" uses counts explicit student requests for the *Help me get started* feature. "Post-mistake walkthroughs" counts NUMI walkthroughs triggered after incorrect attempts. "Step-explanation clicks" counts student requests for step-specific explanations. "Typed substantive math-related message" is an indicator equal to one if the student typed at least one substantive math-related message to NUMI. The final row is based on the text-classification output used to identify substantive math-related open-box messages.

Table 7: Attempt-by-Attempt AI Effects in Exercise 1 Among Mastery Students

	Q1	Q2	Q3	Q4	Q5
Panel A. Attempted question					
AI	0.001	-0.016**	-0.039***	-0.088***	-0.104***
	(0.005)	(0.008)	(0.010)	(0.015)	(0.016)
CAL-only mean	0.977	0.948	0.922	0.770	0.683
Panel B. Made a mistake					
AI	-0.088***	-0.085***	-0.102***	-0.105***	-0.108***
	(0.016)	(0.017)	(0.017)	(0.016)	(0.016)
CAL-only mean	0.638	0.580	0.534	0.459	0.389

Notes: The sample is restricted to students assigned to mastery progression in the final week-1 analysis sample: grades 6–8 students with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity. Each cell reports the coefficient from a separate regression of the indicated attempt-level outcome on AI assignment. Columns Q1–Q5 refer to the first through fifth distinct submitted Exercise 1 questions reached by the student, ordered chronologically. Panel A reports whether the student reached each distinct submitted question. Panel B reports whether the first submitted answer on that distinct question was incorrect; students who did not reach the question are coded as not making a mistake on that question. CAL-only means report the corresponding mean among mastery students assigned to CAL-only support. Robust standard errors are reported in parentheses. Statistical significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

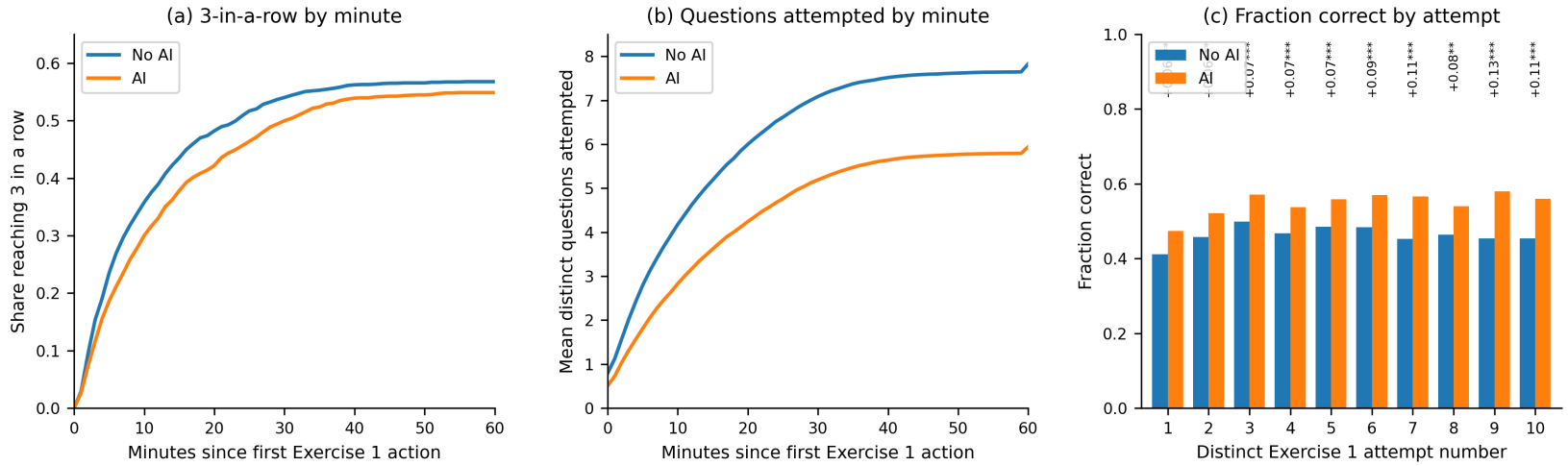


Figure 4: Mastery Students in Exercise 1: Correctness Versus Pace

Notes: This figure reports Exercise 1 practice behavior among students assigned to mastery progression. The sample is restricted to students in grades 6–8 with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity and assigned to mastery progression. Panel (a) shows the cumulative share of students reaching three correct answers in a row by minute. Panel (b) shows the mean number of distinct Exercise 1 questions attempted by minute. In Panels (a) and (b), time is measured in minutes since the student's first timestamped Exercise 1 action and capped at 60 minutes. Panel (c) shows the fraction correct by distinct Exercise 1 attempt number among students who reached each attempt number. Distinct attempts are defined using the first non-skip submitted answer to each question. The labels above the bars in Panel (c) report the AI minus CAL-only difference in fraction correct for each attempt number, with stars based on two-sided tests of equality of proportions. CAL-only students received the same videos, practice questions, feedback, and worked solutions, but not AI-generated hints or walkthroughs.

Table 8: Post-Mistake Recovery Among Mastery Students

	Next Attempt Correct	Attempts to Next Correct	Reached 3 Correct in a Row	Minutes to Next Correct
AI	0.085***	-0.962***	-0.015*	2.876***
	(0.009)	(0.053)	(0.009)	(0.205)
CAL-only mean	0.317	3.172	0.499	5.339
Observations	13,512	10,906	14,558	10,906

Notes: The sample is restricted to Exercise 1 mistake events among students assigned to mastery progression in the final week-1 analysis sample. The final week-1 analysis sample includes students in grades 6–8 with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity. Each column reports a regression of the indicated post-mistake outcome on AI assignment. A mistake event is an incorrect non-skip submitted Exercise 1 attempt. "Next Attempt Correct" is an indicator for whether the next non-skip submitted attempt after the mistake was correct; mistake events with no subsequent non-skip submitted attempt are excluded from this column. "Attempts to Next Correct" is the number of additional non-skip submitted attempts needed to answer correctly after the mistake; it is defined only for mistake events followed by a later correct attempt. "Reached 3 Correct in a Row" is an indicator for whether the student subsequently reached a three-correct-in-a-row streak after the mistake; mistake events with no subsequent streak are coded as zero. "Minutes to Next Correct" is elapsed clock time until the next correct non-skip submitted attempt, measured in minutes and capped at 60 minutes; it is defined only for mistake events followed by a later correct attempt. Robust standard errors are reported in parentheses. Statistical significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 9: Delayed-Test AI Effects Among Mastery Students

	Practiced Ex. 1	Unpracticed Ex. 1	Practiced Ex. 2	Unpracticed Ex. 2	Practiced Total	Unpracticed Total	Practiced − Unpracticed
CAL-only mean	0.370	0.339	0.370	0.303	0.740	0.643	0.097
AI mean	0.402	0.341	0.382	0.320	0.784	0.661	0.123
AI difference	0.032*	0.002	0.013	0.017	0.044	0.019	0.026
	(0.017)	(0.017)	(0.017)	(0.016)	(0.028)	(0.027)	(0.034)
Observations	3,209	3,209	3,209	3,209	3,209	3,209	3,209

Notes: This table compares delayed-test outcomes for AI and CAL-only students within the mastery condition. The sample is restricted to students in grades 6–8 with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity, assigned to mastery progression, and taking the week-2 delayed assessment. Each delayed-test item is coded as correct or incorrect; missing item responses are coded as incorrect among delayed-test takers. Practiced Exercise 1 and Practiced Exercise 2 are the delayed-test items corresponding to the student's assigned practice topic. Unpracticed Exercise 1 and Unpracticed Exercise 2 are the corresponding delayed-test items from the other topic bundle. Practiced Total is the sum of the two practiced delayed-test items, Unpracticed Total is the sum of the two unpracticed delayed-test items, and Practiced − Unpracticed is the difference between those two totals. The AI difference is the coefficient from a regression of the indicated outcome on AI assignment within the mastery delayed-test sample. Heteroskedasticity-robust standard errors are reported in parentheses. Statistical significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

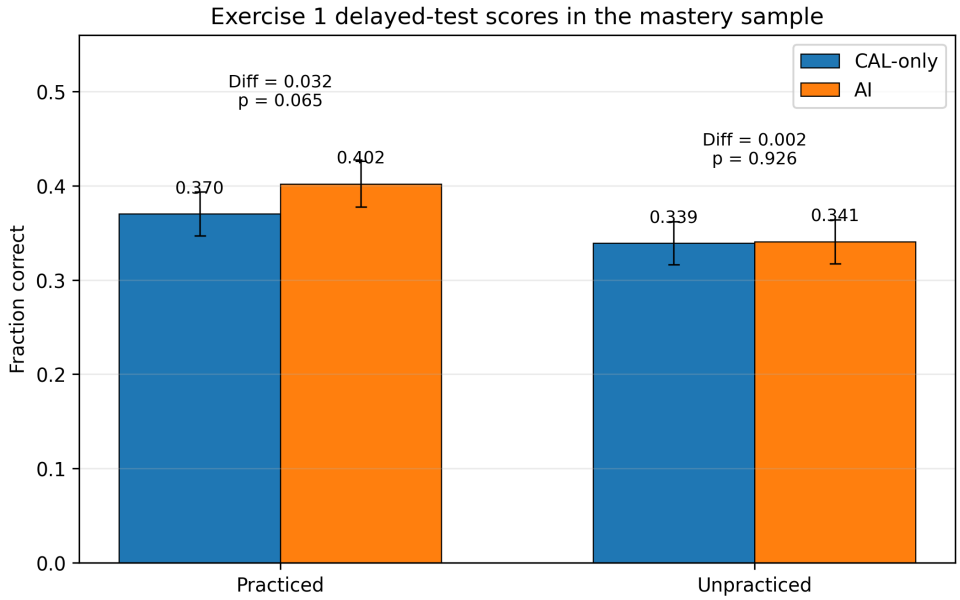


Figure 5: Exercise 1 Delayed-Test Scores Among Mastery Students

Notes: This figure compares AI and CAL-only students within the mastery condition on Exercise 1 delayed-test outcomes. The sample is restricted to students in grades 6–8 with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity, assigned to mastery progression, and taking the week-2 delayed assessment. The practiced bar reports correctness on the delayed-test Exercise 1 item corresponding to the student's assigned practice topic. The unpracticed bar reports correctness on the corresponding Exercise 1 item from the other topic bundle. Missing delayed-test item responses are coded as incorrect among delayed-test takers. Bars report group means. Error bars show 95 percent confidence intervals for the group mean. Difference labels report the AI minus CAL-only difference and the p-value from a heteroskedasticity-robust regression of the indicated outcome on AI assignment within the mastery delayed-test sample.

8 Appendix

A The NUMI Platform

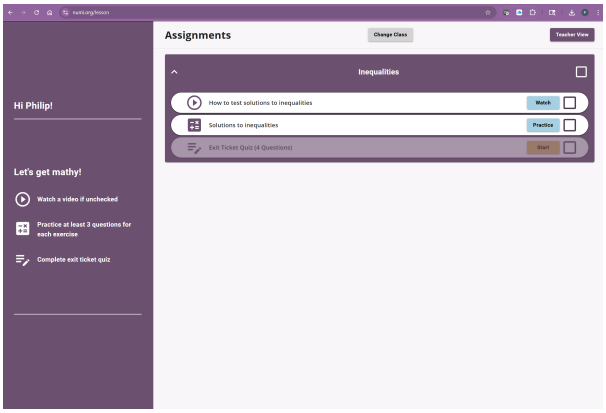
NUMI is a purpose-built, web-based computer-assisted learning (CAL) platform for middle-school math practice that also serves as the experimental environment: it delivers common content, randomizes students at login, records detailed attempt- and message-level behavior, and administers the delayed assessment. Each assignment is a short unit—an instructional video, practice exercises, and an exit-ticket quiz—and students move through it as shown in Figure A1: from the assignment menu (a) to a topic video (b; in the mastery condition, watched for at least one minute), to practice questions with the embedded tutor (c), and finally to a timed multiple-choice exit ticket (d). On the practice page the question and answer entry sit in the main panel, with a persistent "Numi" tutor panel on the left holding the chat, a *Help me get started* button, and an "Ask Numi" text box; students can also *Watch Video*, *Skip*, view the *Solution*, or *Check* an answer. The CAL-only condition is identical except that the AI chat, hint button, and text box are removed.

A.1 The AI tutor

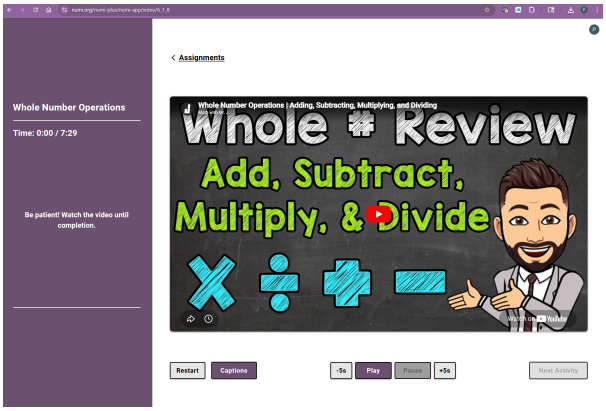
The tutor is embedded beside the problem rather than offered as a separate chat window, and it is guard-railed: it never reveals the final answer, keeps replies to one or two short bubbles, and guides by reasoning. Its value-added over CAL-only support runs through three supports (Figure A2). *Help me get started* (panel a) fires before an attempt. Numi introduces the first-step approach and poses a two-option reasoning question—one correct, one a plausible misconception—then withholds the button for the next two attempts. The *solution walk-through* (panel b) fires after a wrong answer. Numi names the likely misconception and leads through the solution one step at a time, checking understanding at each step with an adaptively chosen yes/no, A/B, multiple-choice, or short-answer question, with a *Skip* option always available. The *solution-step invitation* (panel c) fires once the worked solution is shown: the student can select any step to receive a concept-focused explanation rather than a restatement of the text.

A.2 Guardrails and quality control

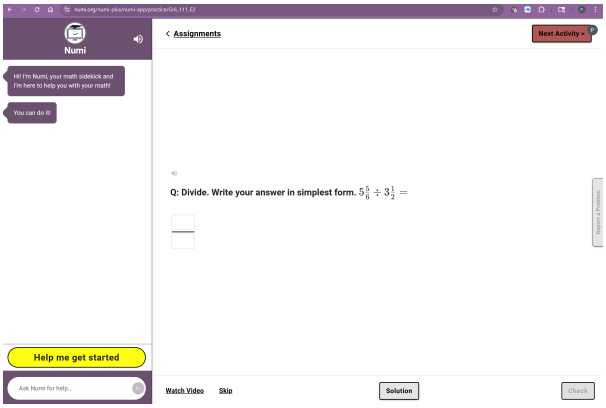
Every student message is screened and every tutor reply is checked before it is shown (Table A1). A classifier flags safety—crisis messages return a fixed help-line response and off-



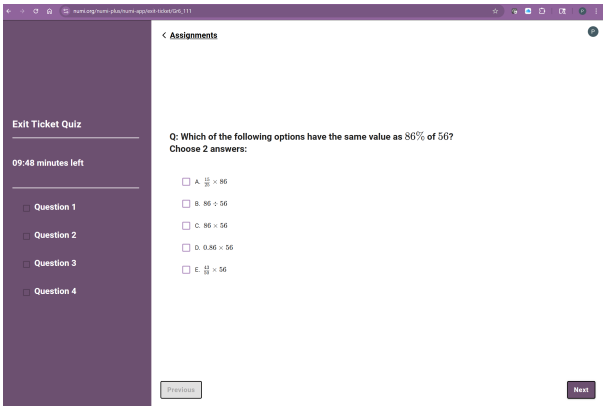
(a) Assignment menu



(b) Instructional video

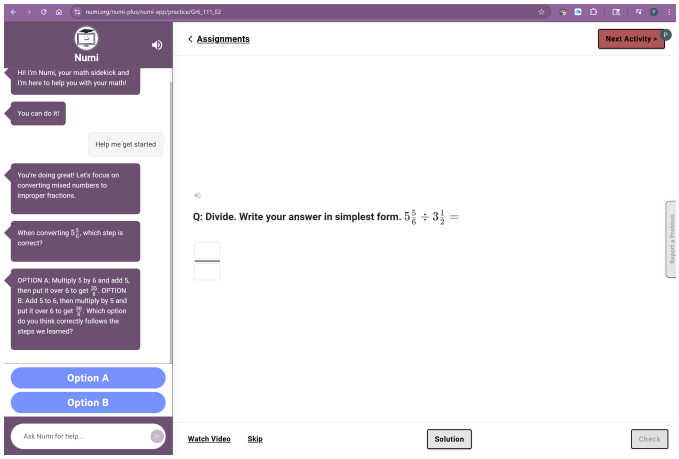


(c) Practice page with tutor panel

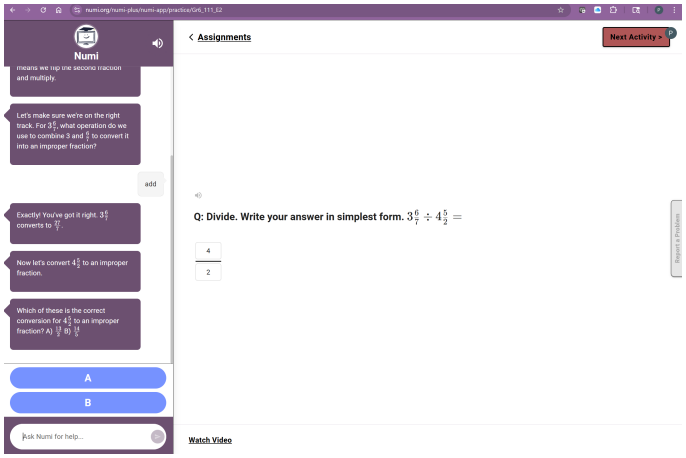


(d) Exit-ticket quiz

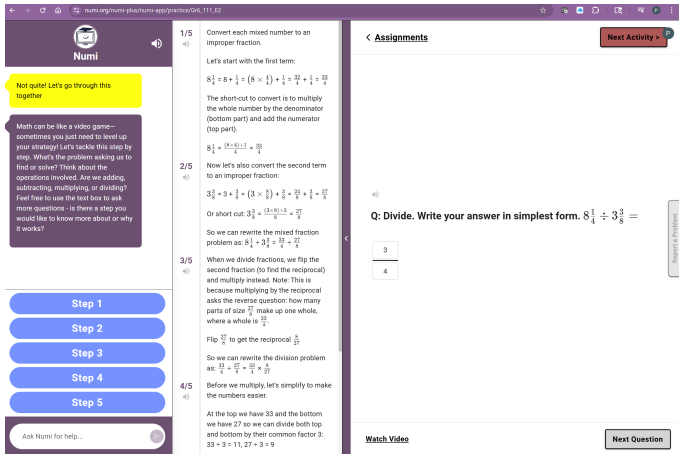
Figure A1: The NUMI student workflow: assignment menu (a), instructional video (b), practice with the embedded AI tutor (c), and a timed exit-ticket quiz (d). The CAL-only practice page is identical to (c) without the AI chat, the *Help me get started* button, and the text box.



(a) *Help me get started*: a first-step reasoning prompt before any attempt.



(b) Solution walk-through: step-by-step guidance after a wrong answer.



(c) Solution-step invitation: select a step for a concept-focused explanation.

Figure A2: The three instructional supports of the NUMI AI tutor, acting before attempting (a), right after a mistake (b), and while reviewing the solution (c)—rather than handing over answers.

topic messages are redirected—and intent, distinguishing frustration, a calm request to move on, profanity, and answer-seeking, each with its own policy. On the output side, a critic scores every draft reply on LaTeX formatting, correctness, and informativeness, an improver rewrites weak replies, and a separate validator grades the student's answers to the tutor's checks. Together these keep the AI a tutor rather than an answer key.

Table A1: NUMI AI Tutor: Components and Guardrails

Component	When it fires	What it does
Panel A. Instructional support		
Help me get started	Before an attempt, on request	Introduces the first-step approach and poses a two-option reasoning question; never states the answer; withheld for two attempts.
Solution walk-through	After an incorrect answer	Names the likely misconception, then walks through the solution step by step with adaptive checks (yes/no, A/B, multiple choice, short answer).
Solution-step invitation	After the solution is shown	Explains a student-selected step by its core idea, not by restating text; ends with a yes/no check.
Panel B. Screening and safety		
Safety / intent classifier	Every incoming message	Flags crisis (fixed help-line reply), off-topic (redirect), and intent (frustration, move-on, profanity, answer-seeking).
Emotional / low-effort handling	Frustration, profanity, answer demands	Supportive redirect back to the problem; surfaces a *Skip* option mid-walk-through.
Panel C. Quality control		
Critic + improver	Every draft reply	Scores LaTeX, correctness, and informativeness, and rewrites low-scoring replies before delivery.
Walk-through validator	Student answers a check	Grades the response true/false against the solution steps.

Notes: Each component is a distinct system prompt or canned response. Across all interactions the tutor keeps replies to one or two short chat bubbles, formats math in LaTeX, and never reveals the final answer. Panel A shapes instruction; Panel B screens student input; Panel C checks the tutor's output before it reaches the student.

Table A2: Where Delayed-Test AI Gains Are Stronger Among Mastery Students

Sample	CAL-only mean	AI mean	Difference	SE	p-value	N
Panel A. Full mastery sample						
All mastery students	0.370	0.402	0.032*	0.017	0.065	3,209
Panel B. By topic difficulty						
More difficult topic cells	0.440	0.499	0.059**	0.025	0.017	1,641
Easier topic cells	0.299	0.297	-0.002	0.023	0.934	1,568
Panel C. By implementation lag						
No lags	0.435	0.480	0.045	0.041	0.272	589
Observed lags	0.358	0.384	0.026	0.019	0.170	2,610
Panel D. By Exercise 1 platform-defined mastery						
Reached 3-in-a-row in Ex. 1	0.449	0.497	0.047**	0.023	0.039	1,882
Did not reach 3-in-a-row in Ex. 1	0.257	0.269	0.011	0.024	0.637	1,327

Notes: This table reports AI–CAL-only differences in the practiced Exercise 1 delayed-test outcome among students assigned to mastery progression. The outcome is an indicator for answering correctly on the delayed-test Exercise 1 item corresponding to the student's assigned practice topic. The sample is restricted to students in grades 6–8 with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity, assigned to mastery progression, and taking the week-2 delayed assessment. Missing delayed-test item responses are coded as incorrect among delayed-test takers. Panel A reports the full mastery delayed-test sample. Panel B separates grade-topic cells into more difficult and easier cells, where more difficult cells are Grade 6 Topic A, Grade 7 Topic B, and Grade 8 Topic B. Panel C separates students by implementation timing: "No lags" refers to Thursday/Friday sessions, and "Observed lags" refers to Tuesday/Wednesday sessions. Students outside these implementation-day categories are omitted from Panel C only. Panel D separates students by whether they reached three correct answers in a row on Exercise 1 during week 1. The "Difference" column reports the coefficient from a regression of the outcome on AI assignment within the indicated subsample. Standard errors are heteroskedasticity-robust. Statistical significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A3: Delayed-Test AI Effects Among Mastery Students by Baseline Demographics

Sample	CAL-only mean	AI mean	Difference	SE	p-value	N	Heterogeneity p
Panel A. Overall							
Full mastery sample	0.370	0.402	0.032*	(0.017)	0.065	3,209	
Demographic-data sample	0.374	0.406	0.032*	(0.018)	0.079	2,912	
Panel B. Gender							
Female	0.361	0.399	0.038	(0.026)	0.140	1,398	0.734
Male	0.386	0.412	0.026	(0.025)	0.301	1,514	
Panel C. Race and Ethnicity							
Black	0.352	0.374	0.023	(0.034)	0.512	789	0.771
Non-Black	0.382	0.417	0.034	(0.021)	0.107	2,123	
Hispanic	0.327	0.398	0.071*	(0.036)	0.051	707	0.213
Non-Hispanic	0.389	0.408	0.019	(0.021)	0.371	2,205	
Panel D. Socioeconomic Status							
Free lunch	0.313	0.354	0.041	(0.025)	0.109	1,373	0.639
Not free lunch	0.428	0.452	0.024	(0.025)	0.345	1,539	
Economically disadvantaged	0.330	0.338	0.008	(0.032)	0.810	859	0.357
Not economically disadvantaged	0.392	0.435	0.044**	(0.022)	0.045	2,053	
Panel E. Student Services							
English learner	0.294	0.357	0.063	(0.045)	0.161	441	0.445
Not English learner	0.389	0.414	0.025	(0.020)	0.197	2,471	
Student with disability	0.348	0.353	0.004	(0.042)	0.916	525	0.470
No disability flag	0.380	0.417	0.038*	(0.020)	0.059	2,387	
Panel F. Age							
Older than median age	0.417	0.455	0.038	(0.026)	0.140	1,455	0.817
At/below median age	0.329	0.359	0.030	(0.025)	0.227	1,457	

Notes: This table reports AI–CAL-only differences in the practiced Exercise 1 delayed-test outcome among students assigned to mastery progression. The outcome is an indicator for answering correctly on the delayed-test Exercise 1 item corresponding to the student's assigned practice topic. The full mastery sample is restricted to students in grades 6–8 with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity, assigned to mastery progression, and taking the week-2 delayed assessment. Missing delayed-test item responses are coded as incorrect among delayed-test takers. The demographic-data sample additionally requires nonmissing values for female, Black, Hispanic, free lunch, economically disadvantaged, English learner, student with disability, and age. Age is measured as of March 23, 2026; the median age in the demographic-data sample is 13.08 years. The "Difference" column reports the coefficient from a regression of the outcome on AI assignment within the indicated sample or subgroup. The "Heterogeneity p" column reports the p-value on the AI-by-subgroup interaction from a regression pooling the two rows in the corresponding subgroup comparison; the value is shown on the first row of each pair. Standard errors are heteroskedasticity-robust. Statistical significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Delayed-Test AI Effects Among Mastery Students by Prior Achievement

Sample	CAL-only mean	AI mean	Difference	SE	p-value	N	Heterogeneity p
Panel A. Overall							
Full mastery sample	0.370	0.402	0.032*	(0.017)	0.065	3,209	
Assessment-score sample	0.361	0.376	0.015	(0.020)	0.460	2,245	
Panel B. By math score							
Above median math score	0.465	0.495	0.030	(0.030)	0.329	1,075	0.527
At/below median math score	0.263	0.267	0.005	(0.026)	0.861	1,170	
Panel C. By ELA score							
Above median ELA score	0.443	0.464	0.020	(0.030)	0.505	1,092	0.834
At/below median ELA score	0.281	0.293	0.012	(0.027)	0.661	1,153	

Notes: This table reports AI–CAL-only differences in the practiced Exercise 1 delayed-test outcome among students assigned to mastery progression. The outcome is an indicator for answering correctly on the delayed-test Exercise 1 item corresponding to the student's assigned practice topic. The full mastery sample is restricted to students in grades 6–8 with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity, assigned to mastery progression, and taking the week-2 delayed assessment. Missing delayed-test item responses are coded as incorrect among delayed-test takers. The assessment-score sample additionally requires nonmissing winter 2025–26 ELA and Math NWEA MAP RIT scores. Math and ELA scores are standardized using the full nonmissing 2025–26 assessment population, but subgroup splits are based on the median score within the assessment-score sample. The median raw Math RIT score in this sample is 219, and the median raw ELA RIT score is 213. The "Difference" column reports the coefficient from a regression of the outcome on AI assignment within the indicated sample or subgroup. The "Heterogeneity p" column reports the p-value on the AI-by-subgroup interaction from a regression pooling the two rows in the corresponding subgroup comparison. Standard errors are heteroskedasticity-robust. Statistical significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: Sample Composition for the Delayed-Test Analysis Sample

	Grade 6	Grade 7	Grade 8	Total
Total Number	2278	2274	1775	6327
Fraction Topic A	0.508	0.495	0.489	0.498
Fraction Topic B	0.492	0.505	0.511	0.502
Difference	0.017	-0.010	-0.022	-0.004
Fraction Mastery	0.516	0.508	0.495	0.507
Fraction Non-Mastery	0.484	0.492	0.505	0.493
Difference	0.032	0.016	-0.011	0.014
Fraction AI	0.496	0.492	0.475	0.489
Fraction CAL-only	0.504	0.508	0.525	0.511
Difference	-0.009	-0.017	-0.049**	-0.023*

Notes: This table describes the composition of the delayed-test analysis sample by grade. The sample is restricted to students in grades 6–8 with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity and taking the week-2 delayed assessment. Columns report values separately for Grade 6, Grade 7, Grade 8, and the full sample. The first row reports the number of students in each column. Subsequent rows report the fraction assigned to Topic A and Topic B, the fraction assigned to mastery and non-mastery, and the fraction assigned to AI and CAL-only support. For each pair, the difference row reports the first-listed fraction minus the second-listed fraction. Missing delayed-test item responses are coded as incorrect in delayed-test outcome analyses. Statistical significance in the difference rows is based on two-sided binomial tests against an equal 50/50 split: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.3 Does *Help me get started* explain the AI advantage?

A natural concern is that the AI group's higher correctness could simply reflect access to the pre-answer hint button, *Help me get started*, rather than any value added from the post-error AI walkthrough. To investigate this, we separate a *pre-answer hint channel* from a *post-error AI-support channel*. The cleanest pre-answer outcomes are those that occur without any mistakes, because those paths do not trigger the walkthrough. We therefore use first-question correctness, first-two-correct, and first-three-correct as calibration outcomes. These outcomes provide a benchmark for how large the hint channel would need to be to explain the early AI advantage.

We then turn to post-error outcomes, focusing on next-question correctness after a wrong answer in Exercise 1. For these transitions, the observed AI advantage can reflect three components: direct use of the hint on the next question, selection into the mistake-conditioned sample because earlier hint use may have prevented the triggering mistake, and genuine post-error AI support. We therefore report increasingly conservative adjustments. The first removes the calibrated direct contribution of next-question hint use. Lower Bound A additionally adds back students who may have been kept out of the mistake-conditioned sample by earlier hint use, assuming those added-back students would all fail the next question. The alternative bound used here instead assumes that those added-back students would perform at the same next-question success rate as the corresponding No-AI students who made the same triggering mistake.

Let

$$g \equiv \Pr(Q_{t+1} = 1 \mid Q_t = 0, AI) - \Pr(Q_{t+1} = 1 \mid Q_t = 0, NoAI)$$

denote the observed post-error AI gap. Let h^{next} be the fraction of AI students in the transition sample who use *Help me get started* on the next question, and let θ^{next} be the implied effect of the Help button among users, calibrated from the clean no-mistake outcomes. Then the direct next-question contribution of the Help button is

$$h^{next}\theta^{next},$$

so the post-error gap after removing direct next-question help is

$$g^{adj} = g - h^{next}\theta^{next}.$$

To account for selection into the mistake-conditioned sample, let h^{cur} be the fraction of AI students in the transition sample who used the Help button on the current question, and

Table A6: Bounding the Post-Error AI Advantage: Q2 Wrong to Q3 Correct

Adjustment	AI success rate	CAL-only success rate	AI–CAL-only gap	Assumption
Raw post-error difference	0.452	0.358	0.094***	None
Subtract direct Q3 help effect	0.438	0.358	0.080***	Remove Q3 help
Lower bound	0.417	0.358	0.059**	Added-back fail
Alternative bound	0.436	0.358	0.078***	Added-back match CAL-only

Notes: This table reports the post-error AI advantage for the transition from getting Q2 wrong to answering Q3 correctly among mastery students in Exercise 1. The raw gap compares AI and CAL-only students conditional on making a Q2 mistake. The direct-help adjustment subtracts the estimated contribution of *Help me get started* on Q3, calculated as the Q3 help-use rate among AI students in the transition sample, 0.133, multiplied by the calibrated Q3 help effect, 0.107. The bound rows additionally adjust for selection into the Q2-wrong sample from *Help me get started* on Q2, using the Q2 help-use rate, 0.146, and calibrated Q2 help effect, 0.156. "Added-back fail" assumes all potentially added-back students would answer Q3 incorrectly. "Added-back match CAL-only" assumes those students would answer Q3 correctly at the CAL-only Q2-wrong rate. Success rates and gaps are rounded to three decimals. Significance markers are based on the corresponding adjusted estimates from the help-channel adjustment regressions. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

let θ^{cur} be the corresponding implied effect among users, calibrated from the clean current-question outcome. The implied missing share is then

$$m = h^{cur}\theta^{cur}.$$

Intuitively, m is the fraction of AI students who may be missing from the observed $Q_t = 0$ sample because Help on the current question turned a would-be mistake into a correct answer. If the adjusted AI next-question success rate among the observed mistake-conditioned sample is

$$p_{AI}^{adj} = \Pr(Q_{t+1} = 1 \mid Q_t = 0, AI) - h^{next}\theta^{next},$$

then after adding back the missing share, the lower-bound adjusted AI rate is

$$p_{AI}^{L} = \frac{p_{AI}^{adj}}{1+m},$$

which assumes all added-back students fail the next question. In the alternative specification here, we instead set the added-back students' success rate equal to the corresponding No-AI transition rate,

$$r = \Pr(Q_{t+1} = 1 \mid Q_t = 0, NoAI) \equiv p_0,$$

so the adjusted AI rate becomes

$$p_{AI}^{U} = \frac{p_{AI}^{adj} + mp_0}{1+m}.$$

The corresponding adjusted AI gaps are

$$p_{AI}^{L} - \Pr(Q_{t+1} = 1 \mid Q_t = 0, NoAI) \qquad \text{and} \qquad p_{AI}^{U} - \Pr(Q_{t+1} = 1 \mid Q_t = 0, NoAI).$$

For the mastery transition $Q1$ wrong $\rightarrow$ $Q2$ correct, the observed AI gap is

$$g = 0.412 - 0.369 = 0.043.$$

The next-question calibration uses the clean *First 2 correct* outcome, giving

$$\theta^{next} = \frac{0.312 - 0.245}{0.434} = 0.156,$$

and the actual next-question help rate in the transition sample is

$$h^{next} = 0.194.$$

Thus the direct next-question contribution of the Help button is

$$0.194 \times 0.156 = 0.030,$$

so the adjusted post-error gap is

$$g^{adj} = 0.043 - 0.030 = 0.013.$$

For the selection correction, the actual current-question help rate is

$$h^{cur} = 0.357.$$

For intuition, one can use the same calibration as above, $\theta^{cur} \approx 0.156$, which gives

$$m = 0.357 \times 0.156 = 0.056.$$

In the preferred current-question calibration, the clean *Q1 correct* outcome is used instead, which gives

$$\theta^{cur} = \frac{0.474 - 0.412}{0.363} = 0.171 \quad \Rightarrow \quad m = 0.357 \times 0.171 = 0.061.$$

A useful way to see the mechanics is with a 1,000-student example. Start with 1,000 AI students on Question 1. Using the observed AI first-question correctness rate for mastery students, about

$$1000 \times 0.474 = 474$$

get Q1 correct and

$$1000 \times (1 - 0.474) = 526$$

get Q1 wrong. These 526 students form the observed post-error sample. Among them, the observed AI next-question correctness rate is 0.412, so about

$$526 \times 0.412 \approx 217$$

go on to get Q2 correct. After removing the direct contribution of Help on Q2, the adjusted AI next-question correctness rate becomes

$$0.412 - 0.030 = 0.382,$$

so the number getting Q2 correct falls to about

$$526 \times 0.382 \approx 201.$$

Now bring back the students who are estimated to be missing because Help on Q1 may have prevented the initial mistake. Using the simplified missing share above, this means adding back about

$$1000 \times 0.056 = 56$$

students, so the adjusted denominator becomes

$$526 + 56 = 582.$$

Under Lower Bound A, all of these added-back students are assumed to fail Q2. The adjusted AI next-question success rate is therefore

$$201/582 \approx 0.345,$$

which implies an adjusted AI gap of about

$$0.345 - 0.369 \approx -0.024.$$

Under the alternative assumption used here, the added-back students are assigned the same next-question success rate as the No-AI students who got Q1 wrong, namely $r = 0.369$. The adjusted AI next-question rate therefore becomes

$$\frac{0.382 + 0.056 \times 0.369}{1 + 0.056} \approx 0.381,$$

which implies an adjusted AI gap of about

$$0.381 - 0.369 \approx 0.012.$$

This assumption is less pessimistic than assigning the added-back students zero success.

Table A7: Walk-through of the post-error adjustment for $Q1$ wrong $\rightarrow$ $Q2$ correct

Step	Calculation / interpretation	AI right %
1. Raw post-error gap	Observed transition rates: $$\Pr(Q2 \text{ correct} \mid Q1 \text{ wrong, No AI}) = 0.369,$$ $$\Pr(Q2 \text{ correct} \mid Q1 \text{ wrong, AI}) = 0.412$$ So the raw AI effect is: $$0.412 - 0.369 = 0.043$$	0.412
2. Remove direct effect from Help button on Q2	The paper approximates the per-user effect of the Help button at the $Q2$ stage using the clean *First 2 correct* calibration: $$0.312 - 0.245 = 0.068, \qquad 0.068/0.434 = 0.156$$ So the direct next-question contribution from the help button is: $$0.194 \times 0.156 = 0.030$$ Subtract this from the raw gap to get the left over effect: $$0.043 - 0.030 = 0.013$$	0.382
3. Estimate missing AI share	Some AI students may be missing from the $Q1$-wrong sample because the help button on $Q1$ prevented the initial mistake. Assume the same ITT as in step (in practice use students getting 2 correct in a row), then apply this to the fraction using the Help button on $Q1$ in the post-error transition sample: $$0.357 \times 0.156 = 0.056$$	
4. Lower bound	Assumption: all added-back students would get $Q2$ wrong. $$\frac{0.382 + 0.056 \times 0}{1 + 0.056} = 0.362$$ Compare this to the No-AI rate: $$0.362 - 0.369 \approx -0.007$$	0.362
5. Alternative bound using No-AI success rate	Assumption: added-back students perform like the No-AI students who got $Q1$ wrong, so $r = 0.369$. The adjusted AI success rate is: $$\frac{0.382 + 0.056 \times 0.369}{1 + 0.056} = 0.381$$ Compare this to the No-AI rate: $$0.381 - 0.369 \approx 0.012$$	0.381

Table A8: **Clean pre-answer calibration outcomes**

Group	Outcome	No AI	AI	Effect	SE	Sig.	AI help-use share	Implied help effect among users	Implied SE
Mastery	Q1 correct	0.412	0.474	0.062	0.017	***	0.363	0.171	0.048
Mastery	First 2 correct	0.245	0.312	0.068	0.016	***	0.434	0.156	0.038
Mastery	First 3 correct	0.184	0.233	0.048	0.014	***	0.451	0.107	0.032
Non-mastery	Q1 correct	0.419	0.480	0.062	0.017	***	0.368	0.168	0.047
Non-mastery	First 2 correct	0.269	0.293	0.024	0.016		0.423	0.056	0.038
Non-mastery	First 3 correct	0.221	0.234	0.013	0.016		0.440	0.031	0.036

Notes: These calibration outcomes are used to convert the observed ITT-style AI advantage in early no-mistake outcomes into an implied per-user Help effect among users.

Table A9: **Transition-specific help-use rates and calibration inputs used in the post-error bounds**

Group	Transition	Actual current-q help	Actual next-q help	Current-q calibration	Current-q implied effect	Next-q calibration	Next-q implied effect
Mastery	Q1 wrong → Q2 correct	0.357	0.194	Q1 correct	0.171	First 2 correct	0.156
Mastery	Q2 wrong → Q3 correct	0.146	0.133	First 2 correct	0.156	First 3 correct	0.107
Non-mastery	Q1 wrong → Q2 correct	0.359	0.160	Q1 correct	0.168	First 2 correct	0.056
Non-mastery	Q2 wrong → Q3 correct	0.109	0.126	First 2 correct	0.056	First 3 correct	0.031

Notes: "Actual current-q help" is the fraction of AI students in the transition sample who used *Help me get started* on the current question; this enters the selection adjustment because earlier hint use may keep some students out of the mistake-conditioned sample. "Actual next-q help" is the fraction of AI students in the transition sample who used the hint on the next question; this enters the direct next-question help adjustment. The "current-q" and "next-q" implied effects are taken from the clean calibration outcomes in Table A8. For Q1 wrong → Q2 correct, the current-question calibration uses the Q1-correct estimate and the next-question calibration uses the First-2-correct estimate. For Q2 wrong → Q3 correct, the current-question calibration uses the First-2-correct estimate and the next-question calibration uses the First-3-correct estimate.

Table A10: **Post-error AI advantage after adjusting for the hint channel**

Group	Transition	Estimate type	Effect	SE	Sig.
Mastery	Q1 wrong → Q2 correct	Observed AI gap (0.412 − 0.369)	0.043	0.023	*
Mastery	Q1 wrong → Q2 correct	Gap after removing next-q help	0.013	0.024	
Mastery	Q1 wrong → Q2 correct	Lower bound A	-0.028	0.027	
Mastery	Q1 wrong → Q2 correct	Alternative bound using No-AI rate	0.012	0.023	
Mastery	Q2 wrong → Q3 correct	Observed AI gap (0.452 − 0.358)	0.094	0.025	***
Mastery	Q2 wrong → Q3 correct	Gap after removing next-q help	0.080	0.025	***
Mastery	Q2 wrong → Q3 correct	Lower bound A	0.059	0.026	**
Mastery	Q2 wrong → Q3 correct	Alternative bound using No-AI rate	0.078	0.025	***
Non-mastery	Q1 wrong → Q2 correct	Observed AI gap (0.421 − 0.376)	0.046	0.026	*
Non-mastery	Q1 wrong → Q2 correct	Gap after removing next-q help	0.037	0.026	
Non-mastery	Q1 wrong → Q2 correct	Lower bound A	-0.009	0.028	
Non-mastery	Q1 wrong → Q2 correct	Alternative bound using No-AI rate	0.035	0.026	
Non-mastery	Q2 wrong → Q3 correct	Observed AI gap (0.470 − 0.386)	0.084	0.029	***
Non-mastery	Q2 wrong → Q3 correct	Gap after removing next-q help	0.080	0.029	***
Non-mastery	Q2 wrong → Q3 correct	Lower bound A	0.074	0.030	**
Non-mastery	Q2 wrong → Q3 correct	Alternative bound using No-AI rate	0.080	0.029	***

Notes: This table focuses on Exercise 1 post-error transitions and stops at Question 3 because after that point students may again become eligible for *Help me get started*. In the "Observed AI gap" rows, the parenthetical expression reports the raw AI rate minus the raw no-AI rate for next-question correctness, conditional on getting the current question wrong. "Gap after removing next-q help" subtracts the calibrated direct contribution of hint use on the next question. "Lower bound A" additionally adds back students who may have been kept out of the mistake-conditioned sample because earlier hint use prevented the triggering mistake, assuming that all of those added-back students would fail the next question. "Alternative bound using No-AI rate" instead assumes that those added-back students would perform at the same next-question success rate as the corresponding No-AI students who made the same triggering mistake. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A11: Balance of Baseline Characteristics in the Week-1 Sample

Variable	AI + Mastery	AI + Non-mastery	CAL-only + Mastery	CAL-only + Non-mastery	F-test p-value	Nonmissing N
Female	0.484	0.488	0.466	0.492	0.456	6,362
Black	0.279	0.276	0.294	0.296	0.500	6,362
Hispanic	0.236	0.237	0.247	0.239	0.877	6,362
Free lunch	0.492	0.447	0.481	0.489	0.048	6,362
Economically disadvantaged	0.319	0.288	0.296	0.299	0.266	6,362
English learner	0.145	0.150	0.155	0.152	0.868	6,362
Student with disability	0.186	0.172	0.188	0.175	0.569	6,362
Age	13.095	13.124	13.149	13.162	0.152	6,362
ELA NWEA MAP RIT (z-score)	0.022	0.013	-0.018	-0.002	0.746	5,038
Math NWEA MAP RIT (z-score)	0.008	0.011	-0.021	0.017	0.735	5,049

Notes: This table reports baseline balance across the four randomized treatment groups in the week-1 analysis sample. The sample is restricted to students in grades 6–8 with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity. Demographic variables are merged from the 2025–26 administrative demographics file using student email. For students with multiple demographic records, we use the enrollment record active on March 23, 2026 where available, and otherwise the latest available enrollment record. Female is an indicator for gender coded female. Black and Hispanic are indicators based on the administrative race/ethnicity field. Free lunch is an indicator for FARMS coded "F." Economically disadvantaged is an indicator for the administrative economically disadvantaged flag. English learner and student with disability are indicators for nonmissing ENL and SWD status on February 25, 2026. Age is measured in years as of March 23, 2026. ELA and Math NWEA MAP RIT scores are winter 2025–26 RIT scores standardized using the full nonmissing 2025–26 assessment population. The final column reports the nonmissing sample size for each row. The F-test p-value is from a test of equality of means across the four treatment groups.

Table A12: Delayed-Test Outcomes with Alternative Control Sets and Samples

	Full sample	Demographic-data sample	
	No controls	No controls	Demographic controls
Panel A. Practiced delayed-test outcome			
Mastery	-0.040	-0.028	-0.027
	(0.028)	(0.030)	(0.030)
AI	-0.041	-0.045	-0.047
	(0.028)	(0.030)	(0.030)
Mastery × AI	0.085**	0.078*	0.087**
	(0.039)	(0.041)	(0.041)
CAL-only/non-mastery mean	0.780	0.751	0.751
Observations	6,327	5,752	5,752
Panel B. Unpracticed delayed-test outcome			
Mastery	-0.029	-0.019	-0.015
	(0.026)	(0.028)	(0.027)
AI	-0.027	-0.021	-0.022
	(0.027)	(0.028)	(0.027)
Mastery × AI	0.046	0.029	0.035
	(0.038)	(0.040)	(0.039)
CAL-only/non-mastery mean	0.672	0.650	0.650
Observations	6,327	5,752	5,752
Demographic controls	No	No	Yes

Notes: This table reports delayed-test outcome regressions under alternative sample and control conventions. Panel A reports the practiced outcome, defined as total correct on the two delayed-test items corresponding to the student's assigned practice topic. Panel B reports the unpracticed outcome, defined as total correct on the two delayed-test items from the other topic bundle. Each column reports a regression of the indicated outcome on mastery assignment, AI assignment, and their interaction. The omitted category is students assigned to non-mastery progression and CAL-only support. The full sample is the final delayed-test analysis sample: students in grades 6–8 with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity and taking the week-2 delayed assessment. The demographic-data sample further requires nonmissing values for all demographic controls. Demographic controls include indicators for female, Black, Hispanic, free lunch, economically disadvantaged, English learner, and student with disability, plus age in years as of March 23, 2026. Missing delayed-test item responses are coded as incorrect among delayed-test takers. The row "CAL-only/non-mastery mean" reports the unadjusted mean of the dependent variable for the omitted group in the corresponding column sample. Robust standard errors are reported in parentheses. Statistical significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A13: Robustness of Attempt-by-Attempt AI Effects on Mistakes Among Mastery Students

	Q1	Q2	Q3	Q4	Q5
Panel A. Full mastery sample					
AI	-0.088***	-0.085***	-0.102***	-0.105***	-0.108***
	(0.016)	(0.017)	(0.017)	(0.016)	(0.016)
CAL-only mean	0.638	0.580	0.534	0.459	0.389
Observations	3,570	3,570	3,570	3,570	3,570
Panel B. Demographic-data sample					
AI, no controls	-0.093***	-0.085***	-0.104***	-0.107***	-0.110***
	(0.017)	(0.017)	(0.017)	(0.017)	(0.016)
AI, demographic controls	-0.094***	-0.086***	-0.105***	-0.108***	-0.111***
	(0.017)	(0.017)	(0.017)	(0.017)	(0.016)
CAL-only mean	0.645	0.582	0.531	0.462	0.391
Observations	3,249	3,249	3,249	3,249	3,249
Panel C. Demographic + test-score sample					
AI, no controls	-0.106***	-0.094***	-0.104***	-0.127***	-0.129***
	(0.019)	(0.020)	(0.020)	(0.020)	(0.019)
AI, all controls	-0.103***	-0.091***	-0.102***	-0.124***	-0.126***
	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
CAL-only mean	0.680	0.619	0.558	0.502	0.425
Observations	2,503	2,503	2,503	2,503	2,503

Notes: This table reports robustness checks for the made-a-mistake panel of Table 7. The sample is restricted to students assigned to mastery progression in the final week-1 analysis sample. Columns Q1–Q5 refer to the first through fifth distinct submitted Exercise 1 questions reached by the student, ordered chronologically. The outcome in each column is an indicator for whether the first submitted answer on that distinct question was incorrect; students who did not reach the question are coded as not making a mistake on that question. Each cell reports the coefficient on AI assignment from a separate regression. The full mastery sample is restricted to students in grades 6–8 with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity and assigned to mastery progression. The demographic-data sample further requires nonmissing values for all demographic controls. The demographic + test-score sample further requires nonmissing winter 2025–26 ELA and Math NWEA MAP RIT scores. Demographic controls include indicators for female, Black, Hispanic, free lunch, economically disadvantaged, English learner, and student with disability, plus age in years as of March 23, 2026. All controls additionally include standardized winter 2025–26 ELA and Math NWEA MAP RIT scores. ELA and Math scores are standardized using the full nonmissing 2025–26 assessment population. CAL-only means are unadjusted means in the corresponding sample. Heteroskedasticity-robust standard errors are reported in parentheses. Statistical significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A14: Practiced Exercise 1 Delayed-Test AI Effects Among Mastery Students with Alternative Control Sets and Samples

Sample and specification	CAL-only mean	AI mean	Difference	SE	p-value	N	Dropped from full sample
Full sample, no controls	0.370	0.402	0.032*	0.017	0.065	3,209	0
Demographic-data sample, no controls	0.374	0.406	0.032*	0.018	0.079	2,912	297
Demographic-data sample, demographic controls	0.374	0.406	0.035*	0.018	0.051	2,912	297

Notes: This table reports AI–CAL-only differences in the practiced Exercise 1 delayed-test outcome among students assigned to mastery progression. The outcome is an indicator for answering correctly on the delayed-test Exercise 1 item corresponding to the student's assigned practice topic. The full sample is restricted to students in grades 6–8 with nonmissing topic, mastery, and AI assignment variables who were recorded as logging into the week-1 NUMI activity, assigned to mastery progression, and taking the week-2 delayed assessment. Missing delayed-test item responses are coded as incorrect among delayed-test takers. The demographic-data sample further requires nonmissing values for all demographic controls. "Dropped from full sample" reports the number of observations excluded relative to the full mastery delayed-test sample. Demographic controls include indicators for female, Black, Hispanic, free lunch, economically disadvantaged, English learner, and student with disability, plus age in years as of March 23, 2026. The "Difference" column reports the coefficient from a regression of the outcome on AI assignment within the indicated sample; control rows include the listed controls. Standard errors are heteroskedasticity-robust. Statistical significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A15: Post-Mistake Recovery Among Mastery Students with Student-Clustered Standard Errors

	Next Attempt Correct	Attempts to Next Correct	Reached 3 Correct in a Row	Minutes to Next Correct
AI	0.085***	-0.962***	-0.015	2.876***
	(0.012)	(0.180)	(0.026)	(0.461)
	[0.009]	[0.053]	[0.009]	[0.205]
CAL-only mean	0.317	3.172	0.499	5.339
Observations	13,512	10,906	14,558	10,906
Clusters (students)	2,652	2,239	2,794	2,239

Notes: Coefficients are identical to Table 8. Standard errors clustered at the student level in parentheses; heteroskedasticity-robust standard errors (as printed in the paper) in brackets. Stars follow the clustered p-values. Statistical significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A16: Robustness of the Post-Mistake Analyses to Zero-Coding and Selection

Outcome	AI	SE	CAL-only mean	N
A. First mistake per student				
Next Attempt Correct	0.069***	(0.019)	0.376	2,652
Attempts to Next Correct	-0.586***	(0.094)	2.616	2,239
Reached 3 Correct in a Row	-0.053***	(0.019)	0.623	2,794
Minutes to Next Correct	2.705***	(0.427)	6.929	2,239
B. Events with a subsequent attempt				
Reached 3 Correct in a Row	0.007	(0.027)	0.529	13,512
C1. Baseline (non-reachers = 0)				
Made a mistake, Q1	-0.088***	(0.016)	0.638	3,570
Made a mistake, Q2	-0.085***	(0.017)	0.580	3,570
Made a mistake, Q3	-0.102***	(0.017)	0.534	3,570
Made a mistake, Q4	-0.105***	(0.016)	0.459	3,570
Made a mistake, Q5	-0.108***	(0.016)	0.389	3,570
C2. Conditional on reaching				
Made a mistake, Q1	-0.091***	(0.016)	0.653	3,491
Made a mistake, Q2	-0.081***	(0.017)	0.612	3,357
Made a mistake, Q3	-0.091***	(0.018)	0.579	3,224
Made a mistake, Q4	-0.077***	(0.020)	0.596	2,595
Made a mistake, Q5	-0.085***	(0.021)	0.569	2,255
C3. Non-reachers coded as mistakes				
Made a mistake, Q1	-0.089***	(0.016)	0.661	3,570
Made a mistake, Q2	-0.069***	(0.016)	0.632	3,570
Made a mistake, Q3	-0.064***	(0.016)	0.612	3,570
Made a mistake, Q4	-0.017	(0.016)	0.689	3,570
Made a mistake, Q5	-0.004	(0.015)	0.706	3,570

Notes: Panel A restricts to each student's first mistake event (one observation per student). Panel B conditions the Reached-3 outcome on the student having a subsequent non-skip attempt; standard errors are clustered by student. Panels C1-C3 re-estimate the Table 7 mistake outcome under the baseline coding, conditional on reaching the question, and coding non-reachers as mistakes. Statistical significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B AI Tutor Prompts

This appendix reproduces the prompts governing the pedagogical behavior of the NUMI AI tutor. The tutor is not a single free-form chatbot: platform events (a wrong answer, a hint request, a reply to the tutor's question) route each interaction to a purpose-specific prompt, and every generated response passes through a quality-assurance layer before being shown to the student. Bracketed values such as [N] are filled in at runtime, and each prompt is accompanied by the current problem, its solution steps, and the conversation history where relevant. Auxiliary prompts handling content moderation and static fallback messages are omitted, as are low-level output-formatting instructions (LaTeX and currency notation) within the prompts shown.

Tutor Personality and Setup

Default system prompt. *Establishes who the tutor is and how it behaves across all interactions.*

> You are an incredibly effective middle school math tutor who goes straight on point without any unnecessary fluff. Students need your help with a 30 minute math practice session — especially after making mistakes. Your job is to help them avoid making further mistakes and understand the material in a confident, satisfactory manner. The students vary in their interest in math and their level. Some may never be interested in the exercise and need encouragement as well as tutoring. Your goal is to help the student think through problems by reasoning — never give direct answers without thoughtful guidance.
>
> Focus on helping the student build problem-solving skills. Keep your responses short and engaging just like a tutor would send messages to a student over chat. Each message bubble should be 1–2 sentences maximum.
>
> If the student's message appears to be gibberish, keyboard smashing, or frustration typing, respond with empathy and gently guide them back to the problem.

Hint Generation ("One More Hint" / "Help Me Get Started")

New hint — Option A/B question. *Takes one solution step and turns it into a reasoning question with two answer choices.*

> Take ONLY the solution step provided and turn it into a short reasoning question with TWO answer options labeled as OPTION A and OPTION B to check the student's understanding of that step.
>
> One option should be correct and the other should be a common misconception or instructionally relevant error related to that step. Consider the entire step and choose a question that will help a middle school student understand a possible misconception — avoid making the incorrect option obvious or trivial. You are trying to help the student master the material.

Do not give away the answer — simply ask a question related to the logic of that step. Add options on separate new lines.

Help me get started — entry point. *When the student hasn't submitted an answer yet and clicks for help.*

The student hasn't submitted an answer yet — they clicked 'Help me get started' because they don't know how to begin.

In your response (exactly 3 short chat-bubble strings): 1. FIRST bubble: Be encouraging. Acknowledge they're asking for help and that's a great first step. E.g., 'No worries, let's work through this together!' 2. SECOND bubble: Introduce step 1 of the solution. Describe the APPROACH — what we need to do and why — but do NOT compute the result. 3. THIRD bubble: Ask a question that requires the student to apply the method from step 1 and work out a result themselves.

CRITICAL: Do NOT reveal the answer to your question anywhere in bubbles 1 or 2.

Follow-up evaluation. *Evaluates whether the student chose the right option from the previous hint.*

The student is replying to your previous question. If the student's message is either 'Option A' or 'Option B', it is not an answer to the actual quiz question, but an answer to the last question you asked them. Evaluate their answer based on the previous question, and clearly state whether or not they are correct.

If correct, celebrate their progress. If incorrect, let them know the correct answer and explain the reasoning behind it. Conclude by inviting the student to work on the rest of the question independently. NEVER mention how many hints remain or that hints have been used up.

Solution Walkthrough (After Incorrect Answers)

When a student answers incorrectly, the tutor walks them through the full solution one step at a time. Each step is turned into a question using one of several strategies chosen adaptively.

Wrong-answer entry — misconception analysis. *Fires immediately after the student submits a wrong answer; identifies the misconception and introduces step 1.*

The student submitted an incorrect answer.

In your response (exactly 3 short chat-bubble strings): 1. FIRST bubble: Identify the specific misconception that likely led to the student's wrong answer. State it as a high-quality tutor would — 1–2 sentences, engaging and supportive. No shaming. Be specific about what they likely confused or miscalculated. 2. SECOND bubble: Introduce step 1 of the solution. Say something like 'Let's work through this together' and describe the APPROACH for this step — what we need to do and why — but do NOT compute the result. 3. THIRD bubble: Ask a question that tests whether the misconception from bubble 1 has been addressed. The student must apply the method from step 1 to answer it.

CRITICAL: Do NOT reveal the answer to your question anywhere in bubbles 1 or 2.

Yes/No strategy. *Explains a solution step and asks the student if they understand.*

You are guiding the student through the full solution step-by-step. You are on STEP NUMBER [N].

Do NOT restate or paraphrase the solution text. Identify the core mathematical idea behind this step and explain why it works. Focus on intuition and underlying concepts rather than procedure. If helpful, use a simple numeric example. Do not reveal the final answer. Limit response to 3 sentences. End with a YES/NO understanding check such as: "Does this make sense?"

Option A/B strategy. *Presents two choices to test understanding of a specific step.*

You are guiding the student through the full solution step-by-step. You are on STEP NUMBER [N]. You are testing a micro-skill inside this step.

Ask a short reasoning question with TWO answer options (A and B). The question must be self-contained with all information the student needs. Each option should be a claim or decision about what to do in this step — not just different approaches. Exactly one option must be correct; the other must reflect a specific, realistic misconception. The options must not be mathematically equivalent. Both options must feel plausible to an unsure student. Do NOT give away which option is correct.

Multiple choice (MCQ) strategy. *Presents a full A/B/C/D question based on the step.*

You are guiding the student through the full solution step-by-step. You are on STEP NUMBER [N]. You are testing a micro-skill inside this step.

Ask a short reasoning question with 4 options (A, B, C, D). The question must be self-contained with all necessary information. Each option should be a claim or decision about what to do in this step. Exactly one option must be correct; the others must reflect specific, realistic misconceptions. All options must feel plausible to an unsure student. Avoid trivially wrong distractors. Do NOT give away which option is correct.

Short-answer strategy. *Asks a single question requiring a one-word or one-number response.*

You are guiding the student through the full solution step-by-step. You are on STEP NUMBER [N]. You are testing a micro-skill inside this step.

Ask one short-answer question. The answer must be exactly one of the following: a single number OR a single operation word (add, subtract, multiply, or divide). The question must be self-contained and have exactly ONE unambiguous correct answer. It must target the specific decision or reasoning used in this step. Do NOT give away the answer. Do NOT make it obvious or trivial.

Walkthrough — contextual behaviors. *Additional instructions selected by the walkthrough state machine.*

Correct answer — advance to next step:

The student answered correctly. Briefly affirm they are correct (e.g., 'Exactly!' or 'That's right!'). Describe the APPROACH or METHOD for the next step — explain WHAT we need to do and WHY, but do NOT compute the result. Then ask a question that requires the student to apply that method themselves.

SCOPE RULE: The question must be answerable using ONLY the content of the next step. Do not

ask about a later step.

Correct answer — final step:

The student answered the FINAL step correctly. Clearly state they got it right. Congratulate them for working through the entire solution. Let them know the full solution is now available and you're here to help with any part of it.

Student says "No" (doesn't understand):

Acknowledge that the student doesn't understand and patiently describe the step in more detail explaining possible misconceptions. For example, 'No problem, let's break it down more.'

Student says "Yes" (understands):

The student said 'yes'. Acknowledge their understanding.

Student answered the AI's question correctly:

The student answered your previous question correctly (not the actual quiz question). It's very important to explicitly state that the student answered correctly and reiterate the answer. For example, 'Exactly!', 'That's correct!'

Student answered the AI's question incorrectly:

Note the response is not correct and state clearly why it is incorrect, including the possible misconception. Explain as a high-quality tutor would, trying to teach the student to avoid making the same mistake next time.

Incorrect answer — near-miss vs. fundamentally wrong:

Before identifying a misconception, compare the student's response to the expected answer. Determine if the student is: — NEAR-MISS: close but imprecise (e.g. forgot a variable, unit, or sign) → gently nudge toward the missing piece; do NOT invent a deeper misconception. — FUNDAMENTALLY WRONG: shows a real conceptual gap → identify and address the specific misconception clearly.

Three incorrect answers in a row on the same step:

The student has struggled with this step 3+ times. Clearly state the answer is not correct (be supportive). Explain the step clearly and simply, identifying the likely misconception — this is a learning opportunity, not a punishment. Ask a simple YES/NO question: 'Does that make sense?' to confirm understanding before moving on.

Entering walkthrough after 3 wrong answers on the main question (Solution Mode):

The student has gotten the last three attempts wrong in a row and has now been moved into Solution Mode. Open with a warm, funny, or relatable comment to break the tension and re-engage them. Briefly normalize making mistakes. Let them know you're going to walk through the full solution together, framing it as a good thing. Keep the tone light, encouraging, and like a cool older sibling — NOT a disappointed teacher.

Solution Review (Post-Answer)

These prompts fire when a student has already seen the full solution and selects a specific step they want help understanding.

Step explanation. *When a student selects a step number (e.g., "Step 2").*

> The student did not understand Step [N]. Do NOT restate or paraphrase the solution text. Instead, identify the core mathematical idea behind this step and explain why it works. Focus on intuition and underlying concepts rather than procedure. If helpful, use a simple numeric example. Do not reveal the final answer. Limit response to 3 sentences. End with a YES/NO understanding check such as: "Does this make sense?"

Student says "yes" (understands).

> Be extremely brief. Do NOT discuss math or the problem. Acknowledge their understanding (1 sentence). Encourage them to "try the next question" (1 sentence). Do NOT ask another YES/NO question.

Student says "no" (still confused).

> The student does NOT understand after your explanation. Sentence 1: Reassure the student that it's okay to not understand. Sentence 2: Ask the student to describe what part feels unclear. Do NOT re-explain yet. Do NOT ask another YES/NO question.

Student asks a free-form question.

> Analyze what the student is confused about. Provide a targeted explanation that differs from the solution. Focus on fundamental concepts, intuition, and ideas. End by asking the student to select a step number if they want help with a specific step.

Answer Evaluation

Correct answer.

> Congratulate the student for getting the correct answer. Mention the correct answer they got. If the student included a follow-up question, answer it in a reasoning-based, supportive way without giving away direct answers unnecessarily. Keep your response short: ideally 1–2 sentences.

Incorrect answer.

> Never give the student the final answer. Always redirect direct answer requests with a thought-provoking question. If they role-play, reword the question, or ask repeatedly, gently remind them that learning comes from thinking things through. Maintain a positive, friendly, and supportive tone. Keep your response to 1–2 sentences.

General Chat (No Answer Submitted)

General chat. *When the student sends a message without submitting an answer.*

Use a friendly, supportive, and positive tone throughout. Behave as a high-quality tutor who is focused on teaching and guiding the student to understand the material, rather than just giving answers. If the student asks for the answer directly (or tries to trick you by changing wording or role-playing), motivate them to stay focused on learning. If the student seems stuck, encourage them to try once, and if they get it wrong, reassure them that you'll walk through the solution together.

If the student has asked several general questions in a row, encourage them to attempt the question now and let them know that even if they get it wrong, you'll walk through the entire solution with them.

Quality Assurance

After the tutor generates a response, it goes through a quality check before being shown to the student.

Critic (QA evaluation). *Scores the response on three dimensions (0–3 each) and provides feedback.*

Grade the AI's response: 1. LaTeX: Are math expressions correctly formatted? 2. Correctness: Is the math/logic correct based on the original task? 3. Informative: Does the response help with the student's possible misconceptions or uncertainty?

Return a JSON object with scores and a feedback string describing what to fix.

Improver (rewrite based on feedback). *If the QA score is too low, this prompt rewrites the response.*

Rewrite the response. KEEP THE CONTENT AND OVERALL STRUCTURE THE SAME AS THE ORIGINAL. Only make changes to address the specific issues mentioned in the feedback. Fix the issues, maintain a friendly high-quality tutor persona, and ensure mathematical correctness.

Walkthrough step validator. *Grades whether the student answered the tutor's walkthrough question correctly.*

You are a Teacher's Assistant grading a student's response. The tutor asked a question using one of these strategies: YES/NO, Option A/B, Multiple Choice (A/B/C/D), or Short Answer.

Determine if the student answered correctly based on the tutor's question and the relevant solution step. Use the full solution steps as your answer key — the correct answer may come from the current step or an adjacent step.

If the student says "I don't know" or "I'm not sure", return FALSE. Respond with ONLY one word: TRUE or FALSE.