Research

Video AI Tutoring:
Experimental Evidence That Conversational AI Improves Learning Outcomes

Dr. Kelly Puzio
Director of Research, Creatium

Video AI Tutoring: Experimental Evidence That Conversational AI Improves Learning Outcomes

Kelly Puzio, PhD

Abstract

Tutoring has become one of education's largest interventions, reaching approximately 10% of U.S. public school students at an annual investment of $20 billion USD per year—roughly $4,000 USD per student. Although face-to-face tutoring programs achieve average effect sizes of 0.37 standard deviations, these benefits remain inaccessible to most students due to the prohibitive costs, staffing demands, and logistical complexities. Given these constraints, this study examined whether video-enhanced AI tutoring could serve as a scalable alternative. Using a between-subjects experimental design, 200 participants were randomly assigned to learn with an instructional video (control) or learn with the same video followed by a brief AI tutoring session (treatment). An ANCOVA revealed a significant intervention effect, $F(1, 168) = 9.68$, $p = .002$, with an effect size of $g = 0.48$, exceeding typical face-to-face tutoring impacts. Participants in the AI tutor condition demonstrated 27.7% greater improvement from pretest to posttest (42.93 vs. 33.62 points). Exploratory discourse analysis revealed that the AI tutor consistently employed pedagogical strategies characteristic of effective face-to-face tutoring, including conceptual prompts, immediate feedback, and misconception repair. These patterns suggest that carefully designed AI tutoring systems can approximate the discourse moves of skilled human tutors while delivering learning gains that surpass those typically observed in face-to-face tutoring programs. Such technologies hold promise for making high-quality tutoring accessible to millions of learners who cannot afford traditional services, potentially transforming individualized instruction from a privilege into a broadly available resource.

Background

Face-to-face tutoring represents one of education's most powerful yet resource-intensive interventions. Recent federal data reveal the scale of current efforts: approximately one in ten public school students receives high-dosage tutoring at an annual cost of $4,000 per student, totaling roughly $20 billion nationwide (Soland, 2023). This investment reflects the robust empirical support for tutoring's effectiveness. Specifically, meta-analytic reviews consistently document meaningful impacts, with average effect sizes of 0.37 standard deviations across a wide variety of programs (Kraft et al., 2024; Nickow, Oreopoulos, & Quan, 2020). Effects are particularly pronounced for programs targeting foundational skills, delivered during school hours with high frequency, and implemented by certified teachers or trained paraprofessionals rather than volunteers.

Yet the very features that make tutoring effective also constrain its scalability. Beyond the considerable per-student costs, implementation challenges include coordinating schedules across thousands of tutor-student pairs, managing frequent absences, and maintaining high-quality relationships. These practical barriers have catalyzed interest in technological alternatives that might democratize access to individualized instruction. The emergence of large language models presents a particularly compelling opportunity to reimagine tutoring systems. However, realizing this potential requires careful attention to the pedagogical mechanisms underlying tutoring's effectiveness—the dialogic interactions, immediate feedback cycles, and adaptive scaffolding that characterize skilled human tutoring. Understanding both what makes tutoring work and what makes it difficult to scale is essential for engineering LLM-based systems that can deliver meaningful learning gains while remaining practically viable for widespread deployment.

Since the advent of computer systems, Intelligent Tutoring Systems (ITS) have evolved from simple drill-and-practice programs to adaptive and data-mining platforms that model student knowledge, provide personalized feedback, and simulate aspects of human dialogue. While ITS produce reliable effects, these gains are typically smaller than well-implemented face-to-face tutoring (e.g., Létourneau et al., 2025; Ma, Adesope, Nesbit, & Liu, 2014). Together, these reviews suggest that ITS can support learning, but the effectiveness of these programs varies substantially.

Evidence from recent meta-analyses suggests that large language model (LLM) supported tutoring can improve multiple dimensions of learning, though the magnitude of these benefits varies by outcome and is complicated by critical methodological issues. Deng et al. (2025) found consistent positive impacts on academic performance, motivation, and higher-order thinking, while effects on self-efficacy were generally non-significant. These patterns suggest that LLM-based tutors may help students perform better and feel more engaged, but do not always build confidence in their own ability to succeed. Importantly, Deng et al. also cautioned that many studies failed to clarify whether the LLM was permitted to be used during post-tests, raising the possibility that some reported gains reflect AI-assisted results rather than genuine learning.

The impact of LLM-based tutoring varies substantially by subject area. Wang and Fan (2025) found that skills and competencies courses (e.g., teacher training, clinical reasoning) demonstrated the strongest gains ($g = 0.87$), followed by language and writing outcomes ($g = 0.53$). In contrast, STEM courses showed smaller average gains in performance ($g = 0.31$), suggesting that generative models may not yet fully support the conceptual reasoning and multi-step problem solving central to scientific and mathematical learning. However, their analyses indicated that GPT-powered tutors fostered higher-order thinking more strongly in STEM than in other domains, suggesting that well-designed applications could leverage LLMs to support scientific understanding. These findings suggest that science-focused LLM tutoring remains an area of both challenge and opportunity. Specifically, science learning demands integration of verbal explanations with visual representations, mathematical reasoning with conceptual understanding, and abstract principles with concrete phenomena.

The incorporation of video into AI tutoring systems represents a promising direction that could address known limitations of text-based instruction. Research on video-based learning demonstrates benefits across engagement, retention, and knowledge transfer. Studies indicate that video resources can increase engagement by 50-60% compared to traditional text materials (Sablić et al., 2020), with students showing 25% better test performance and 30% improved retention when content is properly segmented (Delen et al., 2014; Mayer, 2020). The Cognitive Theory of Multimedia Learning suggests that combining visual and verbal channels can reduce cognitive load while promoting deeper processing (Mayer, 2009), which is particularly relevant for science, where understanding often requires coordinating multiple representations. However, the presence of a hyper-realistic avatar during the video tutoring may, in fact, increase cognitive load and could be distracting for some users.

When designing a virtual tutoring experience, it is critical to understand why face-to-face tutoring works when it does. The mechanisms underpinning the high-quality tutoring align with core learning science principles. Effective tutors engage learners in dialogue that promotes deeper conceptual understanding and metacognition (Chi, 2009), provide immediate feedback (Graesser, D'Mello, & Cade, 2011), target key misconceptions (Hattie & Timperley, 2007), and responsively adapt instruction within each learner's zone of proximal development (Wood, Bruner, & Ross, 1976). Although their implementation may differ across modalities, these principles apply regardless of whether the tutor is human or AI. In addition, tutors form positive social relationships with their students and help students see themselves as capable in this topic or discipline.

Likewise, it is important to consider what is unlikely to work. Cognitive load and multimedia principles clarify why guardrails matter in LLM-based tutoring. When generative systems provide lengthy, poorly segmented explanations, they are likely to overload students' working memory. By contrast, chunked steps, signaling, and targeted hints can conserve cognitive resources and focus the attention of learners. This logic is consistent with early evidence in this area: structured LLM tutors embedded in course material help, while unconstrained LLM tutors can actually reduce learning (Bastani et al., 2025). For designers of LLM-based tutoring systems, the core challenge lies in translating this evidence and principles into a tutoring experience that maintains (or improves on) the benefits of personalized tutoring while avoiding the pitfalls of unstructured, generative AI.

Despite this growing body of evidence, significant gaps remain in understanding how to design and engineer LLM tutoring systems that effectively support learning. Most existing studies have failed to control for AI assistance during assessment and have not strategically implemented principles from the learning sciences. Additionally, no prior studies have examined video-enhanced LLM tutoring systems in science education. This study addresses these gaps by implementing a video-enhanced LLM tutoring experience that incorporates established learning science principles while maintaining rigorous assessment conditions. By examining whether AI tutors can achieve meaningful learning gains in science, this research contributes to our understanding of how to leverage both generative AI and tutoring to support learning at scale.

Study Design and Participants

This study employed a between-subjects design with 200 participants randomly assigned to one of two conditions: a video-only control or a video and AI-tutor treatment. The confirmatory analysis tested whether the AI tutor improved posttest performance relative to the control, while the exploratory analysis examined tutoring transcripts to describe the discourse strategies the AI tutor used. All participants were recruited from Prolific and agreed to participate in a research study. All participants were English-speaking adults living in the US, and all but two reported being born in the US. Data from 25 participants were excluded from the final analytic sample due to incomplete assessment data, failed attention checks, or self-reports of incomplete participation. Table 1 presents the demographic characteristics of participants, by condition:

Table 1
Participant Characteristics, by Condition

Characteristic	**Control** (n = 85)	**AI Tutor** (n = 86)
Age (Mean, SD)	44.06 (12.01)	41.91 (12.58)
Sex (% Female)	59%	58%
Ethnicity (% White)	68%	64%
Education (% Graduate Degrees)	41%	42%
Employment (% Full-Time)	52%	56%

The two experimental groups were comparable across demographic variables. Participants in both conditions were predominantly female, White, and employed full-time, with a mean age of approximately 42 to 44 years.

Task and Conditions

In the control group, participants completed a pretest on tides, viewed a five-minute instructional video, and then completed a posttest. Tides were selected as the focal scientific domain because adults have some background knowledge on this topic but are also likely to hold common misconceptions. The instructional video combined narration, text, and diagrams to explain three core concepts: how the Moon's gravity generates tidal cycles, why two tidal bulges occur on opposite sides of Earth, and how organisms adapt to survive in the intertidal zone. The pretest and posttest were designed to assess participants' understanding of these concepts. In the experimental group, participants completed the same sequence of pretest, instructional video, and posttest, but also engaged in a brief interactive tutoring session after the video and prior to the posttest. The AI tutor was a hyper-realistic avatar powered by a large language model and guided by a proprietary algorithm developed by Creatium. During the session, the AI tutor explained key ideas, answered participant questions, and prompted participants to reason about tidal processes. The tutor was engineered to approximate the talk moves of a skilled human tutor.

Data Collection

All study procedures were administered through a secure online survey platform (SurveyMonkey). Participants first completed the pretest on tides, followed by the assigned instructional materials, and then the posttest. The assessments consisted of multiple-choice questions adapted from established science education instruments and pilot-tested with over 500 adults to ensure reliability and validity. These assessments targeted knowledge of tidal mechanisms, tidal patterns, dual bulges, and intertidal adaptations. The instructional video was embedded within the survey experience. In the AI tutor condition, participants accessed the interactive tutoring session through the survey platform. The tutoring system generated transcripts of these tutoring interactions. To ensure compliance and data quality, attention checks were embedded throughout the survey, and participants who failed these checks, did not complete all assessments, or self-reported incomplete participation were excluded from the analytic sample. This process produced a multimodal dataset that included structured assessment responses, time data, and conversational transcripts.

Data Analysis

The study employed a mixed-methods analytic strategy with two components: a confirmatory impact analysis and an exploratory description of tutoring conversations. To evaluate the impact of the AI tutor, we first examined baseline differences between conditions. An independent-samples *t*-test with Welch's correction was used to compare pretest scores between control and treatment groups, and Hedges' *g* was computed to quantify the standardized mean difference at baseline in alignment with What Works Clearinghouse (WWC) guidelines. To estimate the treatment impact, we conducted an analysis of covariance (ANCOVA) with posttest as the dependent variable, condition as the independent variable, and pretest as a covariate. Adjusted means for each group were estimated at the grand mean of the pretest, and Hedges' *g* was reported as the intervention effect size.

To explore the tutoring interaction, we analyzed transcripts from tutor-student dialogue. Each transcript was coded for the presence of talk moves associated with effective tutoring: metacognitive and conceptual prompts (e.g., "why" or "explain" questions), immediate feedback (confirmations or corrections given directly after learner responses), misconception repair (analogies or explanations provided following weak or inaccurate responses), and adaptive responsiveness (step-by-step guidance, hints, or simplified explanations). For each participant, the proportion of tutor turns containing each talk move was calculated; descriptive statistics characterized how the AI tutor engaged with learners.

Results

The control group scored slightly higher on the pretest ($M = 45.44$, $SD = 18.69$) than the treatment group ($M = 42.30$, $SD = 17.70$). This difference was not statistically significant, $t(169) = 1.13$, $p = .26$, Hedges' $g = 0.17$, indicating no difference at baseline between conditions. Descriptive statistics for pretest, posttest, and time-on-task are presented in Table 2.

Table 2
Descriptive Statistics, by Condition

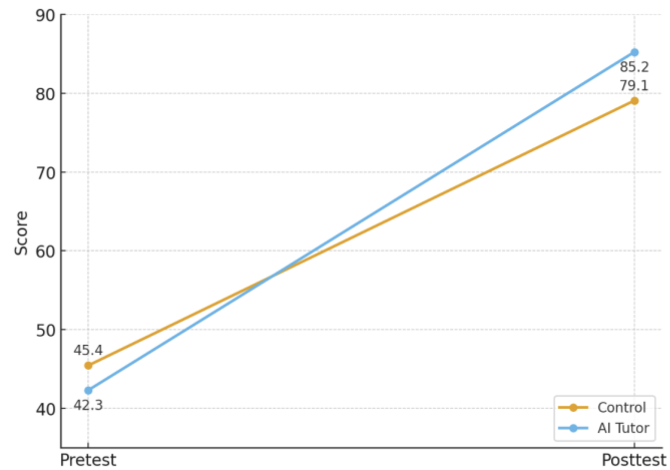
Variable	Control ($n = 85$)	AI Tutor ($n = 86$)
Pretest	45.44 (18.69)	42.30 (17.70)
Posttest	79.06 (16.16)	85.23 (15.24)
Time (minutes)	17.78 (7.36)	26.41 (9.04)

AI Tutor Impact

An ANCOVA with posttest as the dependent variable, condition as the independent variable, and pretest as a covariate indicated a significant treatment effect, $F(1, 168) = 9.68$, $p = .002$. Adjusted posttest means were 78.60 for the control group and 85.69 for the treatment group, resulting in a Hedges' *g* effect size of 0.48, 95% CI [0.17, 0.79]. Gain score analyses showed that the AI tutor group improved by an average of 42.93 points from pretest to posttest compared to 33.62 points in the control group, a 27.7% greater improvement. Together, these findings indicate that participants tutored by the AI tutor performed significantly better on the posttest than controls, even after adjusting for baseline differences. A visualization of this is provided in Figure 1.

Figure 1

Comparison of Pretest–Posttest Performance Across Conditions



Note. Mean scores are shown for participants in both conditions. Both groups improved from pretest to posttest, but the AI tutor group demonstrated a larger gain (42.93 points) compared to the control group (33.62 points), reflecting a 27.7% greater improvement.

Analysis of Tutor Dialogue

Guided by the scientific evidence in this area, every tutor talk turn was coded on seven categories: conceptual press, misconception repair, feedback, adaptive response, explanation, social, and transitions. Examples and descriptive statistics are provided in Table 3.

Table 3
Proportion of Tutor Talk Moves, by Category

Categories	Mean	SD	Example
Conceptual	0.50	0.14	"Why does the moon's gravity lead to two high tides each day?"
Misconception	0.35	0.14	"Not exactly; imagine squeezing a balloon—the sides bulge—this is similar to the two high tides."
Feedback	0.29	0.13	"That's correct—nice work."
Adaptive	0.26	0.11	"Let's break it down: first consider the moon's pull; next, think about Earth's rotation."
Explanation	0.25	0.15	"The moon's gravity pulls ocean water, causing a tidal bulge; as Earth rotates, we get two highs."
Social	0.18	0.13	"Thanks for working through that with me."
Transition	0.05	0.09	"Now let's move on to …"

Note. Values are mean proportions of tutor turns per learner. Because speech turns could be coded in multiple categories simultaneously, proportions sum to greater than 1.

Across these conversations, the AI tutor displayed consistent use of tutoring strategies. Half of all tutor speech turns contained conceptual prompts—questions that pushed learners to explain their reasoning or justify their answers (M = 0.50). In addition, the AI tutor provided immediate feedback, confirming correct responses or addressing errors, in roughly three out of ten turns (M = 0.29). When students struggled, the tutor engaged in misconception repair—offering new explanations or helpful analogies—about 35% of the time (M = 0.35). Adaptive scaffolding, such as breaking problems into smaller steps or providing hints, appeared in approximately one quarter of interactions (M = 0.26). Direct explanations of content occurred with similar frequency (M = 0.25). Beyond these instructional moves, the AI tutor maintained conversational flow through social elements, such as greetings, thanks, and polite acknowledgments (M = 0.18), while using explicit transitions to signal topic shifts in 5% of exchanges (M = 0.05). This pattern provides evidence that the AI tutor balanced challenging students to think deeply, providing support, and maintaining the natural rhythm of a conversation.

Limitations

Several limitations qualify the conclusions of this study. First, the sample consisted of U.S. adults recruited from Prolific, which is important to recognize when considering the generalizability of these findings. Second, the tutoring session was brief and administered immediately after the instructional video, which is a unique instructional context. Third, this study does not tell us how AI tutoring compares to professional human tutoring. These constraints suggest that the present results should be interpreted as an initial demonstration of feasibility and promise rather than as definitive evidence of broad effectiveness.

Discussion

This study provides evidence that a video-enhanced AI tutor can improve learning in science when compared to video instruction alone. The impact ($g = 0.48$) is notable because it exceeds the average impacts reported in meta-analytic reviews of tutoring (Kraft et al., 2024; Nickow et al., 2020). Gain score analyses reinforce this conclusion: learners in the AI tutor condition improved by an average of 42.93 points from pretest to posttest compared to 33.62 points in the control condition, a 27.7% greater improvement. This provides early evidence that LLM-based video tutoring, when carefully designed, can deliver benefits on par with or greater than human-delivered models. This finding reinforces the potential of AI tutoring to extend the reach of effective instructional support without the financial and logistical challenges of scaling face-to-face human tutoring.

The descriptive analysis of transcripts illustrates how these effects may have been realized. The AI tutor consistently employed conceptual prompts, immediate feedback, and misconception repair—three moves emphasized in the research literature as central to effective tutoring (Chi, 2009; Graesser et al., 2011; Hattie & Timperley, 2007). Adaptive responsiveness, though less frequent, indicated that the system attempted to provide instructional scaffolding. These patterns suggest that the tutoring system did more than "present information": it enacted strategies that pressed for reasoning, addressed errors, and supported incremental understanding.

Additionally, the tutor's language included social graces and transitions. Although these accounted for a smaller proportion of turns, they mirror the rapport-building and organizational moves documented in human tutoring. Prior research demonstrates that such relational features are not incidental—they help build rapport, support engagement, and encourage resilience. The presence of politeness markers, brief acknowledgments, and transitions suggests that the tutor was designed not only to deliver content but also to simulate human instructional dialogue. This integration of pedagogical and relational discourse is consistent with emerging evidence that LLM tutors can support motivation and performance (Deng et al., 2025).

The study underscores the value of multimodal design. Consistent with the Cognitive Theory of Multimedia Learning (Mayer, 2009, 2020), video provided a foundation of visual-verbal integration, while the AI tutor reinforced learning through conversational prompts. The complementarity of visual and verbal channels may be particularly promising in STEM domains, where conceptual understanding often depends on linking abstract principles to observable phenomena. While not offered in this tutoring experience, future AI video tutors could provide additional instructional supports, such as interactive simulations, real-time diagram manipulation, and personalized visual scaffolding.

These findings underscore both the promise and caution in scaling AI tutoring. The promise is clear: the intervention produced meaningful learning gains without the costs, scheduling logistics, or staffing demands of face-to-face tutoring. By implementing core pedagogical strategies, AI tutors offer a potential pathway to democratize access to high-quality individualized instruction. The caution, however, is equally important. The single-session format, adult sample, and focus on one science topic constrain our ability to generalize these findings. We do not yet know whether these benefits persist over time, extend to younger learners, or replicate across diverse content areas and student populations. Future research must address these limitations through longitudinal studies that examine sustained use across multiple domains, compare AI tutors directly with skilled human tutors, and test effectiveness with diverse student populations in authentic classroom contexts.

Conclusion

This study demonstrates that a video-enhanced AI tutor can improve learning by embodying conversational and pedagogical strategies long associated with effective tutoring. The AI tutor frequently pressed learners to explain their reasoning, provided immediate feedback, repaired misconceptions, and offered scaffolding—all while integrating social graces that approximate human tutoring. These design features produced measurable gains in conceptual understanding. For researchers, the study highlights the importance of examining how learning science principles can be operationalized in generative AI tutors. For practitioners and policymakers, the results suggest that AI tutoring may provide a cost-effective complement to instruction where access to skilled human tutors is limited. Future work should extend these findings by testing durability, adaptation to diverse learners, and integration into classroom settings. In doing so, the field can continue moving toward the longstanding aspiration of tutoring at scale: substantial gains in learning delivered in ways that are both feasible and equitable.

References

Bastani, H., Bayati, M., Kirschner, P. A., Rahimi, H., & Xu, J. (2025). Generative AI without guardrails can harm learning: Evidence from a randomized controlled trial. *Proceedings of the National Academy of Sciences, 122*(26), e2422633122.

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

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

Delen, E., Liew, J., & Willson, V. (2014). Effects of interactivity and instructional scaffolding on learning: Self-regulation in online video-based environments. *Computers & Education, 78*, 312–320.

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education, 227*, 105224.

Graesser, A. C., D'Mello, S., & Cade, W. (2011). Intelligent tutoring systems. In K. R. Harris, S. Graham, & T. Urdan (Eds.), *APA educational psychology handbook* (Vol. 3, pp. 451–473). American Psychological Association.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Kerac, J., Golubović, G., Milić Keresteš, N., & Ilić, T. (2025). The effects of avatar design on e-learning: A review. In N. Martins & D. Brandão (Eds.), *Advances in design and digital communication V* (pp. 673–686). Springer.

Kraft, M. A., Schueler, B. E., & Falken, G. (2024). What impacts should we expect from tutoring at scale? Exploring meta-analytic generalizability (EdWorkingPaper: 24-1031). Annenberg Institute at Brown University.

Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Léger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems in K-12 education. *NPJ Science of Learning, 10*, 29.

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106*(4), 901–918.

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press.

Mayer, R. E. (2020). *Multimedia learning* (3rd ed.). Cambridge University Press.

Nickow, A., Oreopoulos, P., & Quan, V. (2020). The impressive effects of tutoring: Evidence from randomized experiments (NBER Working Paper No. 27476). National Bureau of Economic Research.

Sablić, M., Mirosavljević, A., & Škugor, A. (2020). Video-based learning (VBL)—Past, present, and future: An overview of the research published from 2008 to 2019. *Technology, Knowledge and Learning, 26*, 1061–1077.

Soland, J. (2023, February 13). Proof points: New federal survey estimates one out of 10 public school students get high-dosage tutoring. *The Hechinger Report.*

Wang, J., & Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications, 12*, 621.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100.

Author's Background

Kelly Puzio, Ph.D., is the Director of Research at Creatium. He is a learning scientist whose work focuses on personalized learning and artificial intelligence. He earned a Ph.D. in Learning, Teaching, and Diversity from Vanderbilt University, a Master's in Education from DePaul University, and a bachelor's degree from the University of Notre Dame. He has been a certified teacher in New Zealand and the United States and brings experience as a faculty member in the Department of Teaching and Learning at Washington State University. Dr. Puzio has served as Principal Investigator on more than $5 million in funding from the National Institutes of Health and the National Science Foundation, and he has been recognized as a Young Scholar by the International Literacy Association.