AI Tutoring Enhances Student Learning Without Crowding Out Reading Effort

Mira Fischer* Holger A. Rau† Rainer Michael Rilke‡

December 22, 2025

Abstract

We study how AI tutoring affects learning in higher education through a randomized experiment with 334 university students preparing for an incentivized exam. Students either received only textbook material, restricted access to an AI tutor requiring initial independent reading, or unrestricted access throughout the study period. AI tutor access raises test performance by 0.23 standard deviations relative to control. Surprisingly, unrestricted access significantly outperforms restricted access by 0.21 standard deviations, contradicting concerns about premature AI reliance. Behavioral analysis reveals that unrestricted access fosters gradual integration of AI support, while restricted access induces intensive bursts of prompting that disrupt learning flow. Benefits are heterogeneous: AI tutors prove most effective for students with lower baseline knowledge and stronger self-regulation skills, suggesting that seamless AI integration enhances learning when students can strategically combine independent study with targeted support.

Keywords: AI Tutors, Large Language Models, Self-regulated Learning, Higher Education
JEL Classification: C91, I21, D83

*BiB - Federal Institute for Population Research, Germany, WZB - Berlin Social Science Center, Germany, and IZA - Institute of Labor Economics, Germany, mira.fischer@wzb.eu.

†Georg-August-Universität Göttingen, Germany, holger.rau@uni-goettingen.de.

‡WHU - Otto Beisheim School of Management, Economics Group, Germany, rainer.rilke@whu.edu.
We thank Paul Herman, co-founder and CPO of the AI learning platform acemate, for collaboration. Financial support by Deutsche Forschungsgemeinschaft through CRC TRR 190 (project number 280092119) is gratefully acknowledged.

1 Introduction

The rapid emergence of generative artificial intelligence (AI) tools since 2022 has had a significant impact on how students learn and how educators teach. In higher education, students increasingly rely on large language models (LLMs) and related tools to study course materials, prepare for assessments, and navigate difficult concepts (Digital Education Council, 2024). These technologies offer unprecedented access to on-demand explanations and examples, raising hopes of scalable, personalized instruction. Emerging evidence suggests that AI-assisted learning may improve educational outcomes (Wu and Yu, 2024). However, its optimal integration into the learning process remains unclear. A critical question is how AI-assisted learning should be designed to foster engagement with learning materials and skill acquisition.

This paper provides causal evidence on the effective design of AI-assisted learning. We focus on the role of LLM-powered AI tutors in university-level exam preparation and experimentally test whether different modes of access to these tutors during study sessions lead to differential outcomes on unaided assessments. Our motivation is rooted in theories of self-regulated learning (e.g., Zimmerman, 2002) and instructional scaffolding (e.g., Wood et al., 1976), suggesting that external support tools—such as AI tutors—can improve performance when they facilitate engagement with learning materials. However, if such tools are used to penetrate a text right from the beginning and avoid cognitive effort they may undermine active information processing and limit the formation of knowledge and understanding. Our experimental design thus introduces contraints on access to an AI tutor in order to examine whether limiting its use to complement students' reading efforts — rather than allowing the AI tutor to be used to substitute for it — improves students' learning outcomes.

Recent work increasingly examines how the design and context of AI usage shape student learning. Bastani et al. (2025) find that unguided access to OpenAI's GPT-4 interface during practice harms unaided exam performance in math exams but this negative effect disappears when instructional scaffolds to the GPT-4 system prompt encourage active engagement and not blunt answers to math problems. The results of De Simone et al.

(2025) point in a similar direction. They show that GPT-4 improves learning outcomes (relative to having no access) only when combined with teacher guidance, curriculum alignment, and structured prompts.[1] Yet a critical question remains: Should AI tutors be continuously available, or restricted until after initial independent study? Existing studies do not test how *degree of AI access* affects unaided performance.

We investigate this design question through a pre-registered artefactual field experiment with 334 university students, in collaboration with an AI learning platform deployed at numerous universities in Germany, Switzerland, and the Netherlands. Its AI tutoring system is based on GPT-4 and uses retrieval-augmented generation (RAG) to respond to student queries with answers specific to course materials. Students study economics textbook material during a 25-minute session to prepare for a subsequent test. In the restricted-access treatment, they gain access to the AI tutoring system after 10 minutes. In the unrestricted-access treatment, they have continuous AI tutor support throughout. In the control treatment, students study the textbook material unaided. All students then complete the same incentivized test without access to the textbook or AI tutor.

The results show that students with AI tutor access significantly outperform those without, with an overall effect of 0.23 standard deviations (sd). Surprisingly, this effect is driven primarily by unrestricted access where students score 0.34 sd higher than students in control and 0.21 sd higher than students with restricted access. Analysis of prompting behavior reveals distinct usage patterns: students with unrestricted access gradually engage with the AI tutor during initial textbook reading, with usage increasing over time. In contrast, students with restricted access initiate intensive AI usage immediately upon gaining access, submitting more and longer prompts with usage decreasing over time. Continuous AI availability appears more conducive to learning than the "reading first, AI tutoring second" structure imposed by restricted access, suggesting that AI tutors function best as adaptive scaffolds aligned with learners' self-regulated engagement, and that, on average, are not used to substitute for reading effort to the degree that it

[1]Related studies examine AI in different educational contexts. Bao et al. (2024) study AI as a complete replacement for human instructors in online Go training, finding that AI outperforms humans and eliminates gender disparities. Fryer et al. (2017) document that student interest in AI tutor-mediated language tasks drops sharply after initial exposure, suggesting novelty effects may limit sustained usage.

harms knowledge acquisition. Heterogeneity analyses indicate, however, that learning gains are concentrated among students with lower baseline performance and stronger self-regulation, specifically those reporting lower procrastination tendencies and a lower preference for distraction. AI tutors thus appear most effective for learners who need content support but can maintain sufficient effort while engaging with the technology.

These results contribute to emerging research on AI integration in higher education by providing causal evidence that flexibility of AI access critically shapes learning outcomes. While prior work emphasizes the importance of teacher guidance, our findings suggest that continuous availability—rather than structured delay—better aligns AI support with self-regulated learning.

2 Experimental Design

2.1 Study Setup

The artefactual field experiment examines the impact of AI-assisted learning on academic performance, comparing different approaches to AI integration during exam preparation. The study follows a two-phase design in which participants first prepare for a baseline test (test 1) and then for a main test (test 2). The study materials comprised two excerpts (5 and 7 pages, respectively) from a well-established introductory economics textbook recognized for its clarity and accessibility to students without prior subject knowledge (Varian, 2016). The excerpts were provided as PDFs and chosen to serve as a high-quality didactic benchmark for textbook-based learning, minimizing potential gains in explanatory clarity from the AI tutor and thereby helping us to focus the analysis on its impact on students' learning behavior. Participants receive instructions on the baseline test and begin with a 15-minute learning period, in which they can read the text but cannot take any notes. Without access to the text they then complete a 10-minute 25-item multiple-choice test (test 1) to assess their initial understanding. After the baseline test, they are asked to estimate how many items they answered correctly and how much they enjoyed preparing for and taking the test.

The second phase introduces the content and the incentives for the main test. Participants are randomly assigned to conditions that vary the availability of the AI tutor (details below) during a 25-minute learning period. After this learning period, participants in all treatments complete the same 10-minute 25-item multiple-choice test (test 2). Then they are asked to estimate their performance in the second test and to rate their enjoyment of preparing for and taking the test. Additionally, subjects in all treatments are asked to estimate how long someone should read the text before engaging with an AI tutor to maximize their performance in the test they just took.

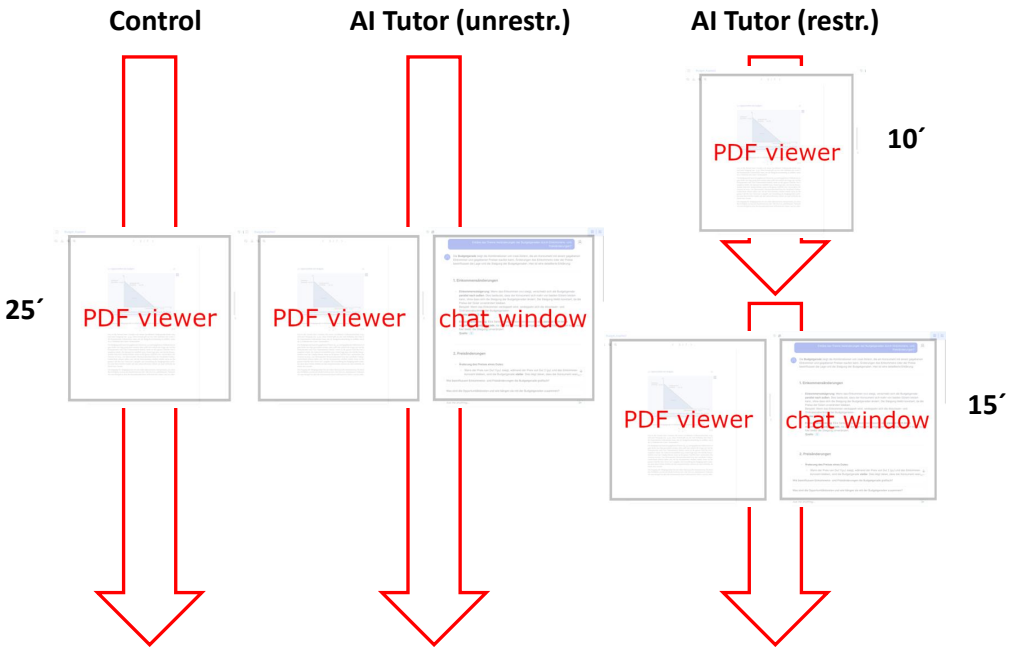
None of the learning or test taking phases can be skipped to finish earlier. Note-taking, annotating or highlighting text is not possible. There are no point deductions for wrong answers. All 50 test items and both estimates of own performance are incentivized with €0.25 each. Participants only learn about their performance on the payoff screen in the end of the experiment.

The AI tutor is implemented as the chatbot of the *acemate* learning platform, a GPT-4-based AI tutoring system. *acemate* is a university-level learning platform designed to transform static course materials into interactive, adaptive learning resources. Central to its functionality is retrieval-augmented generation (RAG) that embeds course content in a vector space and uses semantic similarity search to identify relevant passages when responding to student queries, thereby increasing factual accuracy and ensuring that responses are tailored to the specific instructional context.

2.2 Treatments

In the second phase, participants are randomly assigned to one of three treatment groups, as depicted in Figure 1. In *Control*, participants have no access to the AI tutor and study the textbook excerpt unaided for 25 minutes using a PDF viewer. In treatment *AI Tutor (unrestricted)* participants have full access to both the textbook and AI tutor throughout the entire 25-minute learning phase, with the PDF viewer on the left half of the screen and the chat window on the right half of the screen. In treatment *AI Tutor (restricted)* participants first read the textbook in a PDF viewer for 10 minutes, then transition to

Figure 1: Learning Phase of Main Test



Note: This figure illustrates the learning phase for the main test across the three treatments.

a setup with the viewer on the left and the chat on the right side of their screen for the remaining 15 minutes. In the AI-supported treatments, participants can pose questions freely through the chat interface.

2.3 Sample and Procedures

The experiment was conducted in the experimental economics laboratory of the Technical University of Berlin in February 2025 after receiving ethics approval from the institutional review board of the WZB Berlin Social Science Center. We administered 14 experimental sessions with 24 participants on average (total N=336), each participant sitting in their own cubicle.

The timing of the experiment, instructions, baseline learning, randomization, tests, and survey questions were implemented in Qualtrics. To synchronize the learning phase of the main test across all participants in all treatments, after everyone was done reading their treatment-specific instructions, a two-digit code needed to start the learning phase was publicly announced and had to be entered to start the learning phase. When the study

time was up, experimenters in the lab (one co-author and one or two assistants) ensured and enforced that the AI tutor or PDF viewer window were closed on all computers before a two-digit code needed to transition to the main test was publicly announced.[2] Returning to the AI tutor or the PDF viewer was not possible. Internet access was restricted such that opening any other websites was impossible. Participants were informed that any interaction with the AI chatbot would be recorded and analyzed anonymously. All treatment conditions were represented in each session and participants were individually randomized to one of the treatments, leading to slightly different numbers assigned to each treatment. The final sample included 112 participants in the *Control* Treatment, 102 in the *AI Tutor (unrestricted)* Treatment, and 120 in the *AI Tutor (restricted)* Treatment. A session lasted about 75 minutes with an average pay of 16.72 euros.

3 Hypotheses

The rise of large language models has introduced new tools into the learning environment but it remains unclear whether and how they improve educational outcomes. We focus on one key setting: exam preparation in a university context.

To guide our empirical analysis, we rely on a simple conceptual framework. Students can allocate their limited study time to either reading the assigned text or interacting with the AI tutor. Because chatting with the AI tutor is more engaging and enjoyable for many students than reading, its availability increases overall student effort during the learning period. However, if AI tutor-focussed effort is less effective than effort spent directly reading the text, the net effect of AI tutor availability on learning performance is ambiguous.

Moreover, time spent with the AI tutor is unlikely to be a full substitute for time spent reading. Some initial engagement with the text increases the productivity of subsequent tutor interactions as students need a basic understanding to formulate effective questions and interpret the explanations they receive. Furthermore, students likely not only care

[2]Two participants managed to escape the oversight before the start of the test and were found out during the test to be cheating by keeping the AI tutor or PDF viewer window open in the background. They were recorded and later excluded.

about their academic performance but also their enjoyment during the learning period. As a result, when given unrestricted access, they may tend to use the AI tutor for more time than is optimal for maximizing their learning outcomes.

This framework implies that AI-assisted learning may become more effective if AI tutor access is restricted to induce students to spend more time initially reading the text than they would with unrestricted access.

Below, we formalize this logic into two pre-registered hypotheses.[3] We first test whether access to an AI-based tutor improves learning outcomes relative to studying without AI. Recent studies show that LLM-based tutoring may significantly improve learning outcomes relative to traditional study methods across diverse settings (De Simone et al., 2025; Henkel et al., 2024; Kestin et al., 2024; Vanzo et al., 2025). AI tutors enable more effective learning by providing immediate, adaptive feedback and personalized explanations so that students can clarify concepts as they study and do not have to rely solely on static textbook material (Bastani et al., 2025; De Simone et al., 2025). We therefore expect students with AI tutor access to outperform those studying only with the textbook.

Hypothesis 1 *Participants who have access to an AI tutor will achieve higher performance compared to those in the control group who only study using a text document of a book chapter.*

While access to AI has the potential to improve performance, we expect the way it is integrated into the learning process to be a critical determinant of its effectiveness. In particular, we distinguish between two modes of access: one in which students can rely on AI support throughout the entire study period (unrestricted access), and another in which students study independently before gaining access to AI (restricted access). Our conceptual framework suggests that these differences in mode of access may matter. When students have immediate and unrestricted access to AI, they may use it too soon. This concern is supported by Bastani et al. (2025), who find that unguided AI use can

[3]Our hypotheses were pre-registered on AsPredicted prior to data collection: `https://aspredicted.org/53sy-vhf7.pdf`.

harm learning outcomes when students use AI as a shortcut rather than a learning tool. In contrast, when students are initially required to study independently, they are more likely to engage with the reading material directly. Subsequent AI use then builds on this foundation, allowing students to clarify specific points and reinforce their learning through targeted interaction.

We therefore expect that students with restricted access—those who first engage with the material on their own and only later use AI—will outperform students with unrestricted access.

Hypothesis 2 *Participants in the AI Tutor (restricted) treatment will achieve higher performance compared to those in the AI Tutor (unrestricted) treatment.*

4 Results

To test our hypotheses and investigate treatment effects, we use preregisted OLS regressions of the following form throughout the paper:

$$\text{Test 2 Score}_i = \alpha + \beta \cdot \text{AI Tutor}_i + \gamma \cdot \text{Test 1 Score}_i + X_i \cdot \delta + \varepsilon_i, \tag{1}$$

where α is a common intercept; $AI\ Tutor_i$ is a binary variable equal to 1 if student i was assigned to one of the two AI Tutor treatments, and 0 otherwise; and $Test2Score_i$ is student i's performance in the main test of the experiment. In specifications separately analyzing the effectiveness of the two AI tutor treatments, the AI tutor variable is replaced by two variables: $AI\ Tutor_i$ ($unrestricted$) is equal to 1 if student i was assigned unrestricted AI tutor access for the whole 25 minutes learning period, and 0 otherwise; $AI\ Tutor_i$ ($restricted$) is equal to 1 if student i had access to the text only for the first 10 minutes and gained AI tutor access for the final 15 minutes of the 25 minutes learning period, and 0 otherwise. $Test1Score_i$ is student i's performance in the baseline test of the experiment. X_i is a vector of control variables comprising the field of study (9 categories), the study level (bachelor or master), semester of study (10 categories) and

session fixed effects (14 categories).[4] In this model, β can be interpreted as the average treatment effect.

4.1 Effects of AI Tutor Access on Performance

We begin by testing our pre-registered hypotheses. Our first hypothesis predicted that students with AI tutor access would outperform those without. Column (1) of Table 1 shows that access to an AI tutor during the learning phase significantly improves test performance by 0.23 standard deviations ($p < 0.05$), providing evidence in favor of Hypothesis 1.

Column (2) disaggregates this overall effect by treatment type and allows us to test Hypothesis 2. Students with unrestricted AI tutor access throughout the entire 25-minute learning period significantly outperform control group students by 0.34 sd ($p < 0.01$), while students with restricted access did not perform significantly better than the control group. Contrary to Hypothesis 2, unrestricted access outperforms restricted access by 0.21 sd. This difference is marginally significant (F-test $p = 0.066$).[5] We therefore reject Hypothesis 2. To shed light on why unrestricted access outperforms restricted access—contrary to our prediction—we turn to an analysis of student behavior during the learning phase in the next sub-section.

4.2 Behavior under Restricted vs. Unrestricted AI Tutor Access

The superior performance of unrestricted access challenges our hypothesis that students benefit from forced initial engagement with the textbook. To understand this unexpected result, we examine how students in each treatment allocated their study time and engaged with available learning resources. Our expectation that academic performance would be superior with restricted AI access compared to unrestricted was based on the assumption

[4] As 18% of our sample do not have a German Abitur and 31% do not have an Abitur math grade, we are not controlling for these variables in the main specifications.

[5] In relative terms, AI tutor access overall resulted in a 5.8% higher performance, with restricted access improving performance by 3.4% (not sign.) and unrestricted access by 8.6% relative to the control group. Unrestricted access resulted in a 5.1% higher performance than restricted access.

Table 1: Effects of AI Tutor on Performance

	(1)	(2)
	Test 2 Score	
AI Tutor (any)	0.227**	
	(0.106)	
AI Tutor (unrestricted)		0.337***
		(0.116)
AI Tutor (restricted)		0.132
		(0.122)
Test 1 Score	0.468***	0.459***
	(0.050)	(0.051)
Bachelor	-0.157	-0.158
	(0.101)	(0.101)
Constant	-0.800*	-0.868*
	(0.452)	(0.462)
Field of Study FE	Yes	Yes
Semester FE	Yes	Yes
Session FE	Yes	Yes
p-value (F-test: restr. = unrestr.)		0.066
R^2	0.379	0.385
N (Observations)	334	334

Note: This table shows OLS regressions with the standardized test performance *Test 2 Score* as dependent variable. *AI Tutor (any)* equals 1 if participants had any access to the AI tutor, and 0 if they only had access to the text. *AI Tutor (unrestricted)* equals 1 if participants had full access to the AI tutor, while *AI Tutor (restricted)* equals 1 if access was limited. *Test 1 Score* is the standardized score in the first test. *Bachelor* equals 1 if the student is enrolled in a Bachelor programme, 0 otherwise. *Field of Study FE*, *Semester FE*, and *Session FE* denote fixed effects for field of study, semester, and experimental session, respectively. Heteroskedasticity robust standard errors are in parentheses. * $p < .1$, ** $p < .05$, *** $p < .01$

that students prematurely rely on the AI tutor's explanations, spending minimal time with the textbook.

However, Table 2 reveals a different pattern. Column (1) compares initial reading time before submitting the first prompt across all three treatments. Columns (2)-(4) focus on the two AI treatments to examine prompting behavior: total number of prompts submitted to the AI tutor, prompting intensity (prompts per minute of AI access), and average prompt length in characters.

Column (1) shows that both AI treatments significantly reduced reading time relative to control students, who could do nothing but read for the 25-minute learning phase. Restricted-access students initially read for approximately 11.2 minutes (coefficient: -13.476, $p < 0.001$) before submitting their first prompt, while unrestricted-access students read for 5.7 minutes (coefficient: -18.915, $p < 0.001$). Crucially, unrestricted-access students still spent substantial time reading before engaging with the AI tutor—contradicting concerns about immediate abandonment of independent study.

Column (2) examines total number of prompts. Despite having 10 additional minutes of AI access, unrestricted-access students tended to write slightly fewer prompts overall than restricted-access students (6.1 vs. 7.0), though this difference is not statistically significant (coefficient: -0.836, $p > 0.10$). However, Column (3) reveals a significant difference in prompting intensity: unrestricted-access students submitted 0.23 fewer prompts per minute of AI access ($p< 0.001$). Column (4) shows that unrestricted-access students also wrote prompts that were significantly shorter by on average 9 characters ($p < 0.05$).[6] Taken together, these findings suggest that unrestricted access facilitated a more moderate AI tutor use—students submitted shorter prompts at lower intensity.

Figure 2 reveals the temporal dynamics underlying these aggregate differences in prompting behavior. The bar chart depicts restricted-access (unrestricted-access) students in red bars (hollow blue bars). The spike at 10 minutes shows that restricted-access students begin prompting intensively immediately upon gaining AI access, with prompt frequency declining steadily thereafter ($p < 0.001$, OLS linear fit). Unrestricted-access

[6]Note that the model in Column (4) is estimated using robust regression. For explanation, see table note.

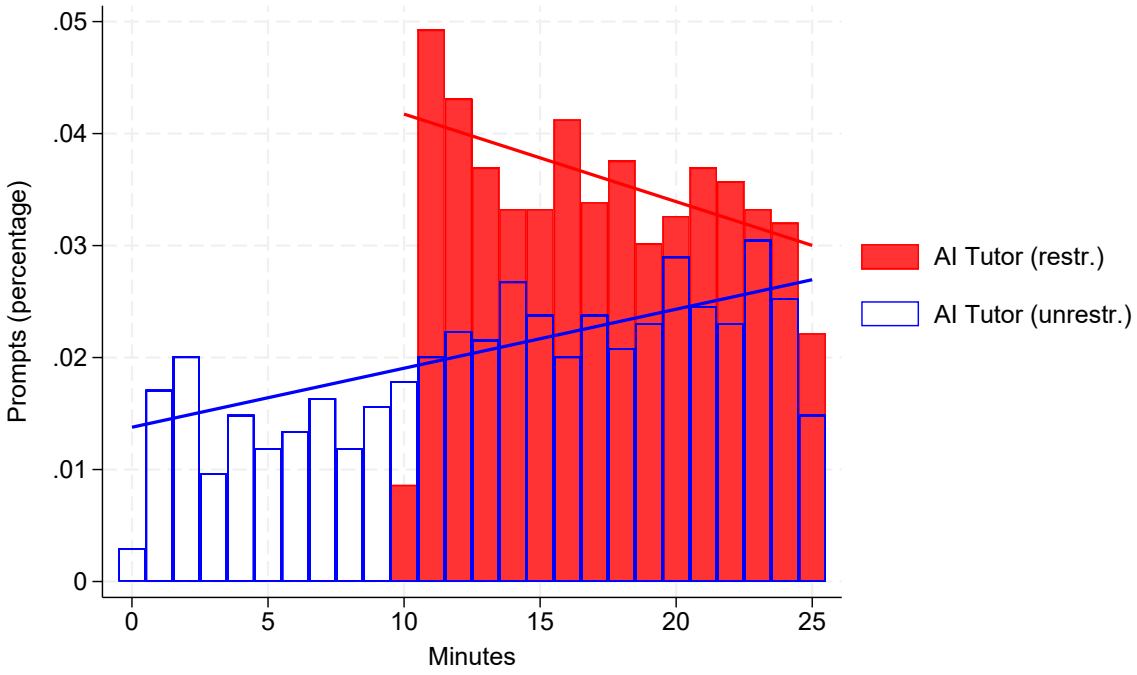
Table 2: Effects of (Un)Restricted AI Tutor on Prompting Behavior

	(1)	(2)	(3)	(4)
		Prompt		
	Reading time	Amount	Per Min.	Length
AI Tutor (unrestricted)	-18.916***	-0.836	-0.234***	-8.864**
	(0.737)	(0.665)	(0.035)	(4.070)
AI Tutor (restricted)	-13.476***			
	(0.378)			
Test 1 Score	0.093	0.037	0.003	0.890
	(0.091)	(0.121)	(0.006)	(0.741)
Bachelor	0.265	0.183	0.008	1.721
	(0.509)	(0.770)	(0.042)	(4.304)
Constant	24.628***	6.945*	0.478***	81.765***
	(2.676)	(3.539)	(0.170)	(25.960)
Field of Study FE	Yes	Yes	Yes	Yes
Semester FE	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	Yes	Yes
R^2	0.792	0.219	0.347	0.131
N	334	222	222	212

Note: This table shows OLS regressions, in columns (1) - (3), analyzing how unrestricted or restricted access to the AI tutor affected participants' prompting behavior during the learning phase. In column (4) robust regression (based on iteratively reweighted least squares) is used to estimate effects on prompt length, which puts less weight on outliers in prompt length (see Figure A. 1 in the Appendix) driven by a few observations in which participants copy-pasted large sections of the learning material into the chat. *AI Tutor (unrestricted)* equals 1 if participants had full access to the AI tutor, while *AI Tutor (restricted)* equals 1 if access was limited. *Reading Time* measures the total time participants spent reading the learning material (in minutes) before they submitted the first prompt. *Prompt Amount* is the total number of prompts sent to the AI tutor. *Prompts per Minute* normalizes the number of prompts by AI tutor access time. *Prompt Length* captures the average character length of participants' submitted prompts. *Test 1 Score* is the standardized score in the first test. *Bachelor* equals 1 if the student is enrolled in a Bachelor programme, 0 otherwise. *Field of Study FE*, *Semester FE*, and *Session FE* denote fixed effects for field of study, semester, and experimental session, respectively. Heteroskedasticity robust standard errors are in parentheses. * $p < .1$, ** $p < .05$, *** $p < .01$

students exhibit the opposite pattern: they gradually increase their prompting over time ($p < 0.001$, OLS linear fit), starting with minimal AI engagement and progressively integrating more support as their understanding develops.

Figure 2: Distribution of Prompts over Time



Note: This figure depicts the distribution of prompts over time. It shows the mean distribution by treatment over all 14 sessions, adjusted for the number of observations per treatment. Times are rounded down to the next minute. Additionally, the linearized trends of prompts over time by treatment are shown.

Unrestricted access thus seems to have enabled scaffolding: students could alternate between reading the text and asking questions to resolve confusion immediately rather than accumulating questions during forced independent study—a flexibility that appears to have enhanced learning effectiveness.[7]

4.3 Heterogeneous Treatment Effects

Table 3 splits the sample at the mean value of of several characteristics to investigate heterogeneous effects of AI tutor access. Each column examines a different moderator: book reading habits (Books, Column 1), baseline subject knowledge (Low Test 1 Score,

[7]In Section A.1 in the Appendix we analyze differences in prompt content between the AI treatments.

Column 2), general procrastination tendencies with learning tasks (Procrastinator, Column 3), preference for distraction (Pleasure, Column 4), and AI chatbot experience (AI Experience, Column 5).

Interestingly, Column (1) shows that stronger book reading habits are associated with significant AI tutor benefits ($p = 0.02$, F-test), while we find no significant effects for weaker reading habits, suggesting that being comfortable with reading books may also make AI-assisted learning more effective. Column (2) indicates that AI tutor benefits are concentrated among students with below-average baseline performance ($p = 0.05$, F-test), suggesting that AI tutors are particularly valuable for students needing additional support. Consistent with the importance of self-regulation, Column (3) shows that AI tutor access significantly improves performance for students with low-procrastination tendencies (0.35 sd, $p < 0.05$), while the effect for stronger procrastinators is not statistically significant ($p = 0.35$, F-test). Similarly, Column (4) reveals significant benefits for students with below-average distraction preference (0.50 sd, $p < 0.001$) but the negative AI×Pleasure interaction eliminates benefits for students with above average preference for distraction ($p = 0.87$, F-test). Column (5) finds no moderating effect of prior AI chatbot experience, as captured by our scales.[8]

Table A. 2 in the Appendix investigates whether the AI tutor benefitted students in some fields of study more than others. The significant treatment effects are concentrated on students in the humanities and in computer science and electronics, who showed the lowest performance in the control group.

These heterogeneity patterns underscore that AI tutors do not improve learning for everyone. Students with stronger self-regulation skills (low procrastination, low distraction seeking) and those with weaker baseline knowledge benefit most, while those prone to distraction or with strong prior knowledge see no significant gains. This aligns with our earlier behavioral findings: effective AI use requires the ability and motivation to integrate support strategically and not excessively substitute direct engagement with the text with AI tutor interaction.

[8]Note that none of these p-values are adjusted for multiple hypothesis testing and that evidence from these heterogeneity analyses should be treated as merely suggestive because of low statistical power.

Table 3: Heterogeneous Effects of AI Tutor on Performance

	(1)	(2)	(3)	(4)	(5)
	Test 2 Score				
AI Tutor	0.129	0.176	0.353**	0.504***	0.258*
	(0.143)	(0.122)	(0.153)	(0.158)	(0.137)
AI×Books	0.215				
	(0.202)				
Books	-0.146				
	(0.169)				
AI×Low Test 1 Score		0.178			
		(0.210)			
Low Test 1 Score		-0.339			
		(0.213)			
AI×Procrastinator			-0.222		
			(0.198)		
Procrastinator			0.118		
			(0.164)		
AI×Pleasure				-0.481**	
				(0.199)	
Pleasure				0.318*	
				(0.167)	
AI×AI Experience					-0.039
					(0.205)
AI Experience					-0.093
					(0.171)
Test 1 Score	0.465***	0.387***	0.469***	0.461***	0.460***
	(0.051)	(0.077)	(0.053)	(0.050)	(0.051)
Bachelor	-0.167*	-0.140	-0.147	-0.163	-0.153
	(0.101)	(0.101)	(0.101)	(0.101)	(0.102)
Constant	-0.728	-0.693	-0.866*	-1.044**	-0.790*
	(0.458)	(0.463)	(0.461)	(0.492)	(0.448)
Field of Study FE	Yes	Yes	Yes	Yes	Yes
Semester FE	Yes	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	Yes	Yes	Yes
p-value (F-test: AI + AI×Ind.)	0.022	0.053	0.351	0.865	0.173
R^2	0.381	0.385	0.382	0.391	0.382
N	334	334	334	334	334

Note: This table shows OLS regressions of participants' standardized test performance *Test 2 Score* on the presence of an AI tutor and individual characteristics. *AI Tutor* is a dummy variable equal to 1 if participants had access to the AI tutor during the learning phase, and 0 otherwise. *Books* is a self-reported measure capturing the frequency of book reading for pleasure. *Low Test 1 Score* is an indicator for participants scoring below the median in Test 1. *Procrastinator* is a self-reported measure indicating the extent to which a participant tends to delay important learning tasks. *Pleasure* captures the importance a participant assigns to not becoming bored during the experiment. *AI Experience* measures prior experience using AI-based tools (mean over three variables capturing experience with AI use for productivity, learning, and pleasure). *Test 1 Score* is the standardized score in the first test. *Bachelor* equals 1 if the student is enrolled in a Bachelor programme, 0 otherwise. *Field of Study FE*, *Semester FE*, and *Session FE* denote fixed effects for field of study, semester, and experimental session, respectively. Heteroskedasticity robust standard errors are in parentheses. * $p < .1$, ** $p < .05$, *** $p < .01$

4.4 Effects on Beliefs and Hedonic Experiences

To better understand the mechanisms underlying our findings and students' interactions with the AI tutor, we examine treatment effects on their beliefs and enjoyment. Column (1) shows that both AI tutor treatments reduce beliefs about optimal reading time before AI use by approximately 1.5 minutes (restricted: -1.50, unrestricted: -1.66; both $p < 0.05$). Specifically, control group students believe 17 minutes is optimal, compared to about 15.5 minutes in both AI treatments. Thus, participants on average held beliefs even more extreme to the one informing our restricted-access treatment, and the 10-minute delay in access we introduced should not have been binding given these beliefs. However, Figure A. 2 in the Appendix shows that while beliefs are similar across AI treatments, access restriction was, indeed, behaviorally binding and induced behavior more aligned with these beliefs—which, however, decreased performance (see Table 1).

Turning to hedonic experiences, Column (2) shows that only restricted AI access significantly increases enjoyment of test preparation relative to control (0.31 sd, $p < 0.05$), while unrestricted access shows no significant effect. Column (3) reveals that neither treatment significantly affects enjoyment of test-taking.

Column (4) examines after-test performance beliefs. Being assigned AI assistance increases confidence by approximately 0.2 sd relative to control (unrestricted: 0.20 sd, $p < 0.10$; restricted: 0.22 sd, $p < 0.05$), with no significant difference between treatments (F-test $p = 0.83$). Comparing these confidence gains to actual performance improvements (unrestricted: 0.34 sd, significant; restricted: 0.13 sd, not significant), unrestricted-access students appear slightly underconfident, while restricted-access students appear somewhat overconfident relative to their gains.

Overall, these findings reveal a paradox: AI tutors boost student confidence and, in restricted access, increase enjoyment and belief-behavior alignment—yet only unrestricted access delivers actual performance gains, suggesting that perceived effectiveness of delayed AI access on average does not align with actual effectiveness.

Table 4: Effects of AI Tutor on Beliefs and Hedonic Experiences

	(1)	(2)	(3)	(4)
	Belief Opt. Reading Time	Enjoyment: Preparation	Enjoyment: Test Taking	Confidence: Test 2 Score
AI Tutor (unrestricted)	-1.661**	0.199	0.043	0.195*
	(0.768)	(0.121)	(0.117)	(0.099)
AI Tutor (restricted)	-1.500**	0.308**	0.183	0.217**
	(0.723)	(0.125)	(0.126)	(0.093)
Test 1 Score	-0.335	0.046	0.110**	0.013
	(0.307)	(0.045)	(0.052)	(0.062)
Bachelor	0.487	-0.115	-0.025	-0.123
	(0.680)	(0.112)	(0.112)	(0.087)
Enjoym.: Test 1 Prep.		0.434***		
		(0.055)		
Enjoym.: Test 1 Taking			0.423***	
			(0.055)	
Confidence: Test 1 Score				0.638***
				(0.047)
Constant	17.271***	-0.364	-0.001	-0.060
	(2.456)	(0.291)	(0.316)	(0.315)
Field of Study FE	Yes	Yes	Yes	Yes
Semester FE	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	Yes	Yes
p-value (F-test: restr. = unrestr.)	0.833	0.388	0.253	0.830
R^2	0.118	0.286	0.313	0.554
N	334	334	334	334

Note: This table shows OLS regressions examining how access to the AI tutor affects participants' beliefs and hedonic experiences. *AI Tutor (unrestricted)* equals 1 if participants had unrestricted access to the AI tutor during the learning phase, while *AI Tutor (restricted)* equals 1 if participants had access under restricted condition. *Belief Opt. Reading Time* is the belief about time im minutes one should spend reading before interacting with the AI tutor to maximize test performance. *Enjoyment: Preparation* and *Enjoyment: Test Taking* capture self-reported enjoyment during the preparation and test-taking phases of Test 2, respectively. *Confidence: Test 2 Score* measures participants' confidence in their performance immediately after test completion (estimated number of correct items). *Test 1 Score* is the participant's score in the first test. *Bachelor* equals 1 if the student is enrolled in a Bachelor programme, 0 otherwise. *Enjoyment: Test 1 Prep*, *Enjoyment: Test 1 Taking* and *Confidence: Test 1 Score* denote baseline measures collected during the first test phase. *Field of Study FE*, *Semester FE*, and *Session FE* denote fixed effects for field of study, semester, and experimental session, respectively. Heteroskedasticity robust standard errors are in parentheses. * $p < .1$, ** $p < .05$, *** $p < .01$.

5 Conclusion

This paper provides causal evidence on how the design of AI tutor assistance shapes students' learning outcomes and study behavior in exam preparation. Using an artefactual field experiment with 334 students, we show that access to an AI tutor increases performance on incentivized assessments by 0.23 standard deviations. These gains are driven by students with unrestricted access, who outperform the control group by 0.34 standard deviations and the restricted-access group by 0.21 standard deviations.

Our results show that when students can flexibly integrate AI support into their learning process, they use it in a measured and productive manner—suggesting that, instead of reducing effort, unrestricted AI access appears to provide instructional scaffolding without crowding out engagement with foundational materials.

These findings challenge a substitution logic in models of effort allocation: if interacting with an AI tutor is more enjoyable than reading, and students care not only about their academic outcomes but also about pleasure, then AI tutor access would lead students to read too little. Yet our behavioral data show the opposite. Students with unrestricted access read for several minutes before incorporating the AI tutor into their learning process. Their prompting behavior unfolds slowly and steadily over time, with shorter prompts. Restricted-access students, by contrast, exhibit a sharp spike in prompting immediately upon gaining access, followed by a decline—consistent with an accumulation of unresolved questions during the enforced reading period, which then triggers a burst of AI engagement.

This pattern illustrates that AI tutoring functions less as a substitute for reading and more as a scaffold that supports comprehension when students experience difficulty. The effectiveness of AI thus depends on allowing students to deploy it autonomously at moments that suit their learning process. A design that imposes a fixed learning sequence, to correct presumed self-control failures, disrupts this self-regulated learning behavior.

Our heterogeneity analyses further highlight the importance of learner characteristics. Students with lower baseline knowledge and stronger self-regulation—particularly those reporting low procrastination tendencies and lower distraction preference—experience the

largest gains. Those with higher distractibility do not benefit from AI access. These patterns are consistent with a behavioral model in which AI tutors raise the productivity of learning effort conditional on the learner's ability to sustain engagement and strategically integrate external support. AI cannot compensate for weak self-regulation but, at least in our setting, neither does it exacerbate it. Instead, AI amplifies the returns to structured, goal-directed learning for students who already possess the behavioral foundations for effective study.

Finally, students' beliefs about optimal study time allocation diverge systematically from the allocation that maximizes performance. Participants generally believe one should read for 16–17 minutes before using an AI tutor, yet unrestricted-access students—who deviate most from this belief—perform best. Restricted-access students on average behave in line with their beliefs but do not achieve better outcomes than the control group. This gap between stated and revealed optimality underscores how learners may hold inaccurate beliefs about how to use AI tutors effectively, and paternalistic constraints that mirror these beliefs may reduce their effectiveness.

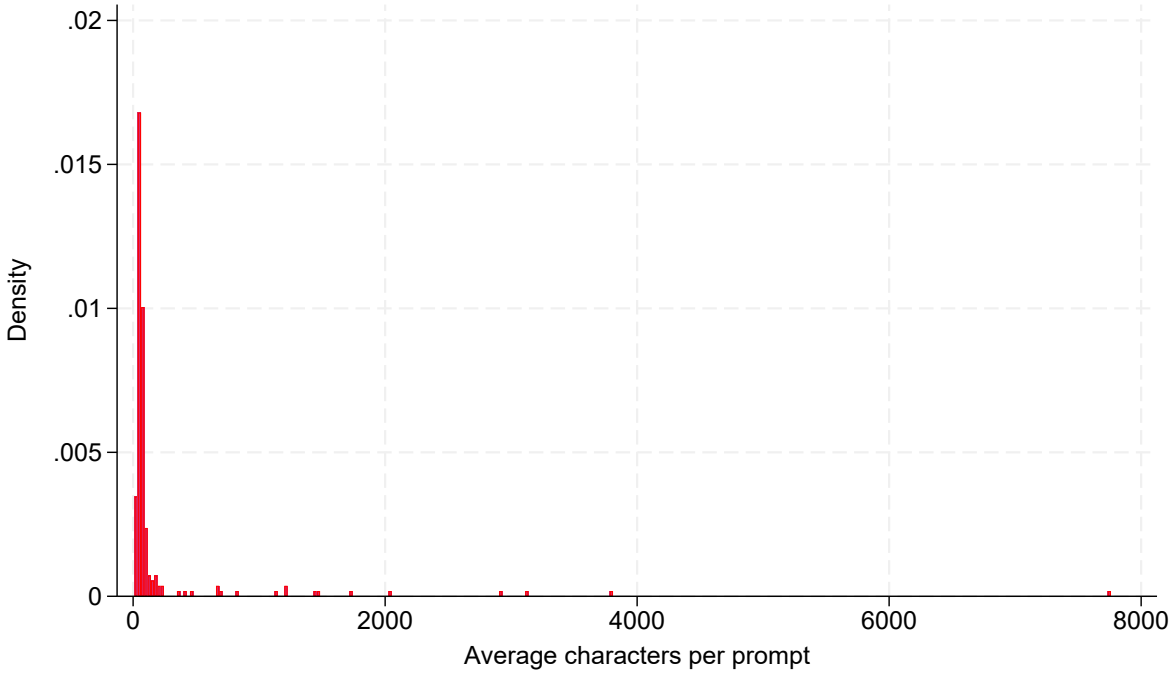
Our findings carry implications for both theory and policy. Our results suggest that many students use AI tutors in ways that reflect sophisticated self-regulation rather than simple effort substitution, and that it is a basic competency needed to use this learning technology effectively. Models of AI-assisted learning may therefore need to incorporate it. They also show that AI tutoring may help to close learning gaps. At the same time, policy-makers considering the large-scale roll-out of AI-assisted learning in education should consider that its benefits may be limited to those students who already possess good learning skills. In order to benefit students across the distribution, investing in their self-regulation ability and learning habits should therefore become an even stronger priority. Teachers should also be trained to support students in effectively incorporating AI tutors into their learning. Yielding to the temptation to save costs by using the technology to (partially) substitute for teachers may exacerbate learning gaps.

Overall, our study shows that in settings where the goal is to promote skill acquisition rather than assess unaided performance, AI tutors can serve as effective complements to

traditional study materials. The long-term effects on knowledge retention, higher-order reasoning, and learning skills remain open questions. Future research should examine repeated or longer-term exposure to AI tutors, their interaction with teaching practices, and their impact on inequality in skill acquisition across different social groups.

A Appendix

Figure A. 1: Distribution of average characters per prompt



Note: This figure shows the distribution of average characters per prompt.

A.1 Prompt Classification

We classified all user-generated prompts using an automated language model–based procedure. The classification scheme was developed inductively from a preliminary review of a random subset of student prompts, in which distinct patterns of learning-related use emerged. Based on this analysis, we defined four main categories: Conceptual Understanding, Applied Problem Solving, Memorization Aids, and Exam Preparation. Each prompt was then submitted to OpenAI's GPT-3.5-Turbo model with a standardized instruction that asked the model to assign the prompt to one of these categories. The model was instructed to select the most relevant category based on the content and intent of the prompt and to avoid ambiguous or multiple labels. This procedure ensured consistent categorization across participants and enabled us to quantify individual differences in how students used the GPT-4 based AI tutor for learning.

Specifically, we used the following prompt to classify user-generated inputs:

```
You are a scientist specializing in education and language analysis. Your
task is to diligently categorize prompts based on how students use ChatGPT
to learn and prepare for a test. Assign the prompt to the most relevant
category. If it does not perfectly match one, choose the closest one.
DO NOT return 'Other' unless absolutely necessary.

Categorize the following prompt:

'<PROMPT>'

Possible categories:

Conceptual Understanding { The student asks for definitions, theoretical
explanations, comparisons, or real-world examples.

Applied Problem Solving { The student asks for solving a mathematical
problem, interpreting a graph, or applying the concept to a numerical
scenario.

Study & Memorization Aids { The student asks for summaries, simplified
explanations, or flashcards.

Exam & Assignment Preparation { The student asks for mock exam questions,
essay structuring help, or proofreading.

Other { If none of the above apply, suggest a new category.

Return only the category name. Do NOT include numbers, descriptions,
or extra text. Return exactly one of the categories listed above.
```

Table A. 1: Effects of Unrestricted AI Tutor on Prompt Content

	(1) Applied Problem Solving	(2) Conceptual Understanding	(3) Exam Preparation	(4) Memorization Aids	(5) Other
AI Tutor (unrestricted)	0.589	-7.619*	-0.280	6.858*	0.452
	(0.876)	(4.249)	(1.895)	(3.476)	(1.395)
Test 1 Score	0.330	-0.292	1.583**	-1.494	-0.127
	(0.244)	(2.332)	(0.783)	(1.983)	(0.613)
Bachelor	1.508*	2.934	-2.968	0.720	-2.193
	(0.822)	(4.612)	(2.389)	(3.665)	(1.479)
Constant	-3.400	89.581***	18.699**	-6.092	1.213
	(2.324)	(15.468)	(9.023)	(10.998)	(4.126)
Field of Study FE	Yes	Yes	Yes	Yes	Yes
Semester FE	Yes	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	Yes	Yes	Yes
R^2	0.153	0.216	0.152	0.138	0.252
N	212	212	212	212	212

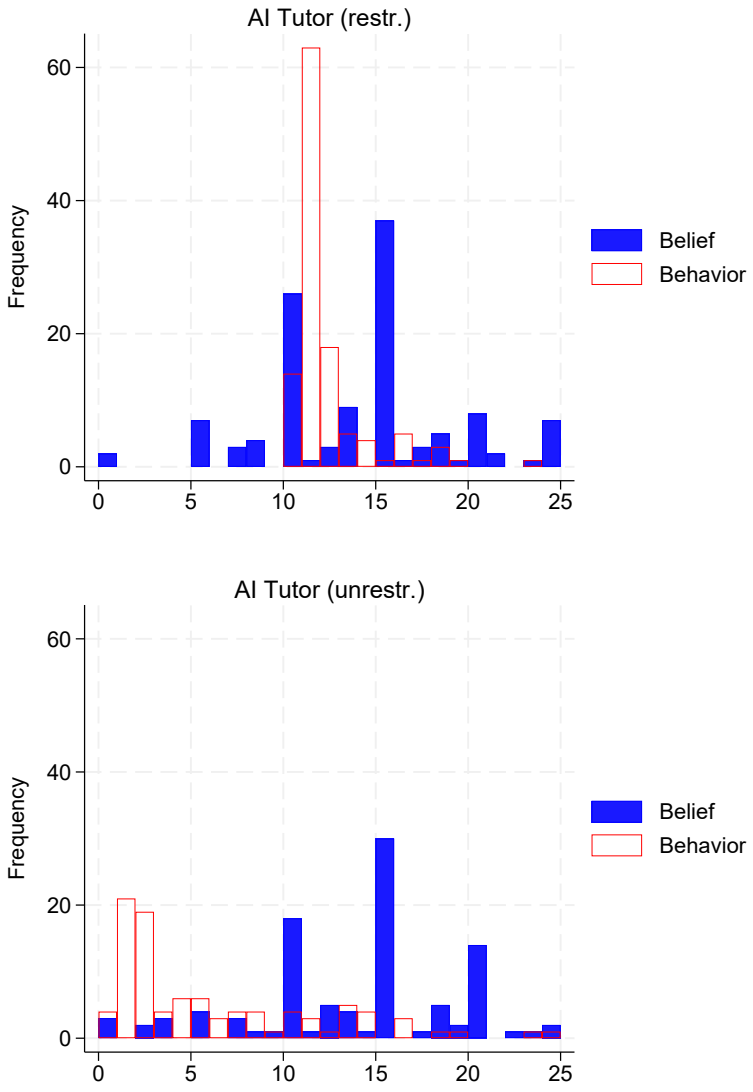
Note: This table reports OLS regressions examining how the use of an unrestricted versus restricted AI tutor affected the thematic content of student prompts. The dependent variable in each column represents a user's share of prompts classified into the respective category. *AI Tutor (unrestricted)* equals 1 if participants had full access to the AI tutor, *Test 1 Score* is the standardized score in the first test. *Bachelor* equals 1 if the student is enrolled in a Bachelor programme, 0 otherwise. *Field of Study FE*, *Semester FE*, and *Session FE* denote fixed effects for field of study, semester, and experimental session, respectively. Heteroskedasticity robust standard errors are in parentheses. * $p < .1$, ** $p < .05$, *** $p < .01$

Table A. 2: Effects of AI Tutor on Performance by Field of Study

	(1) Treatment Effect	(2) Control Group Mean
Humanities	0.593**	−0.797
Maths & Sciences	0.378	−0.073**
Process Engineering	0.123	0.056***
Computer Science & Electronics	0.463*	−0.230**
Transportation & Machine Systems	0.003	0.320***
Construction & Environment	0.153	−0.046
Economics & Management	0.074	0.428***

Note: This table shows OLS regressions with the same specification as in Table (1), Column (1), and adds dummies for the fields of study and their interactions with the treatment variable *AI Tutor (any)*. Column (1) shows AI tutor effects by field (*AI Tutor(any)* + *AI Tutor(any)* $\times$*Field*), Column (2) reports the control group mean, stars indicate significant differences relative to humanities students (reference category). Heteroskedasticity robust standard errors are in parentheses. * $p < .1$, ** $p < .05$, *** $p < .01$

Figure A. 2: Beliefs About Ideal Prompting Behavior versus Actual Behavior



Note: This figure compares how long people actually read (timing of first prompt) with the belief (elicited after behavior) for how long one should read the text before initializing prompting to maximize test performance. The upper panel shows distributions for the restricted AI Tutor treatment, the lower panel shows distributions for the unrestricted AI Tutor treatment. All initial prompts over all sessions are shown. Time of first prompt is rounded down to the next minute.

References

Bao, L., Huang, D., and Lin, C. (2024). "Can artificial intelligence improve gender equality? Evidence from a natural experiment." *Management Science*.

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., and Mariman, R. (2025). "Generative AI without guardrails can harm learning: Evidence from high school mathematics." *Proceedings of the National Academy of Sciences*, *122*(26), e2422633122.

De Simone, M., Tiberti, F., Rodriguez, M. B., Manolio, F., Mosuro, W., and Dikoru, E. J. (2025). "From chalkboards to chatbots: Evaluating the impact of generative AI on learning outcomes in Nigeria." Policy Research Working Paper 11125, World Bank.

Digital Education Council (2024). "Digital Education Council Global AI Student Survey." `https://www.digitaleducationcouncil.com/post/what-students-want-key-results-from-dec-global-ai-student-survey-2024`, accessed: 2025-11-04.

Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., and Sherlock, Z. (2017). "Stimulating and sustaining interest in a language course: An experimental comparison of chatbot and human task partners." *Computers in Human Behavior*, *75*, 461–468.

Henkel, O., Horne-Robinson, H., Kozhakhmetova, N., and Lee, A. (2024). "Effective and scalable math support: Experimental evidence on the impact of an AI-math tutor in Ghana." In *International Conference on Artificial Intelligence in Education*, 373–381, Springer.

Kestin, G., Miller, K., Klales, A., Milbourne, T., and Ponti, G. (2024). "AI tutoring outperforms active learning." *Working Paper*.

Vanzo, A., Chowdhury, S. P., and Sachan, M. (2025). "Gpt-4 as a homework tutor can improve student engagement and learning outcomes." In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 31119–31136.

Varian, H. R. (2016). *Intermediate Microeconomics with Calculus: A Modern Approach: Ninth International Student Edition*. WW Norton & Company.

Wood, D., Bruner, J. S., and Ross, G. (1976). "The role of tutoring in problem solving."

Journal of Child Psychology and Psychiatry, *17*(2), 89–100.

Wu, R., and Yu, Z. (2024). "Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis." *British Journal of Educational Technology*, *55*(1), 10–33.

Zimmerman, B. J. (2002). "Becoming a self-regulated learner: An overview." *Theory into practice*, *41*(2), 64–70.