Experimental Evidence on the Learning Impact of Generative AI*

Zara Contractor Germán Reyes

March 2026

Abstract

We estimate the effects of generative AI on learning through a randomized experiment. Undergraduate students learn a new topic with or without AI access, and we measure outcomes both immediately and one week later. AI access improves test scores by 0.25 standard deviations, with gains concentrated among middle-performing students and persisting one week later. AI access also changed how students write: treated students produced longer essays with shorter sentences and simpler vocabulary that independent graders rated as higher-quality. Learning benefits were partly driven by a reallocation of time from writing toward information-seeking activities.

*Contractor: Middlebury College (zcontractor@middlebury.edu), Reyes: Middlebury College (greyes@middlebury.edu). For helpful discussions and comments, we thank Jeff Carpenter, Amy Collier, Guy Ishai, Cory Koedel, Seunghoon Lee, Peter Matthews, David Munro, Caitlin Myers, Andrea Robbett, and participants in the Cornell behavioral economics group. Financial support from the Middlebury College's Office of the Provost is gratefully acknowledged. We thank Kimberly Barros, Rebecca Deranian, Brooke Dolan, Alice Gindin, Sarah Hayward, Ember Pikramenos, Kevin Ramirez, Ephraim Shinko, and Juan Marcelo Verdugo for their assistance with experiment proctoring. The survey was reviewed and approved by the Institutional Review Board at Middlebury College.

1 Introduction

Every week, over 900 million people use ChatGPT, with college-aged adults being the most active demographic and over a quarter of their messages related to learning (OpenAI, 2025). Given these magnitudes, it is fundamental to understand how generative AI tools affect learning. On one hand, these tools could enhance learning by acting as an on-demand tutor that explains difficult concepts and provides personalized feedback, but also undermine it by substituting for the cognitive effort that learning requires—writing entire essays, completing problem sets, and generating code with minimal student input.

We estimate the immediate and longer-term effects of AI access on learning through a randomized experiment with 210 undergraduate students. The experiment consists of two in-person sessions conducted approximately one week apart. In the first session, participants are randomly assigned to an AI-allowed or an AI-forbidden condition. During a 35-minute learning phase, students learn about a randomly assigned topic and write an analytical essay; the AI-allowed group receives a dedicated ChatGPT (GPT-4o) account and may use any generative AI tools, while the AI-forbidden group uses traditional resources. We monitor AI usage and compliance through direct proctor observation, ChatGPT conversation logs, and self-reports.

We measure learning outcomes along multiple dimensions. All participants complete knowledge tests before and after the learning phase without access to AI or external resources. Essays are independently evaluated by human and AI graders blind to treatment condition across six quality dimensions, and we analyze writing style, textual similarity, and AI-generated content. In the second session, approximately one week later, all participants complete a longer retention test and write a new essay, both without access to any external resources, including AI.

We first document that our random assignment generated substantial variation in AI usage. About 69 percent of students in the AI-allowed condition used generative AI during the learning phase, compared to near-zero usage in the control group. Students primarily used AI for conceptual support rather than text generation: 44 percent used it to explain difficult concepts, 23 percent to summarize readings, and only 16 percent to draft responses. Among AI users, 88 percent reported that AI was helpful for the learning task.

Our main finding is that AI access improved learning outcomes. Students with AI access scored 0.25 standard deviations higher on the immediate knowledge test, a 6.3 percentage point increase in the share of questions answered correctly (on a baseline of 56.3 percent).

This improvement was concentrated among students in the middle of the performance distribution, with smaller effects at the extremes. The gains persisted approximately one week later, though with some attenuation: students who had used AI during initial learning scored 0.27 standard deviations higher on the retention test, about 110 percent of the initial effect.

We also find that AI access changed how students write. Treated students produced essays that were about 7 percent longer, with shorter sentences and simpler vocabulary, and these essays were rated higher-quality by independent graders. Commercial AI content detectors flag a significantly higher fraction of treated students' text as AI-generated, though AI-allowed essays were no more "homogeneous" than control essays, contrasting with the textual convergence documented in workplace settings (Brynjolfsson et al., 2025). These writing changes partially persisted one week later, when students wrote without AI access, consistent with the persistence of learning effects.

To examine mechanisms, we analyze how AI access affected student behavior during the learning phase. AI access did not change total time spent on the task or engagement with provided materials, but it reallocated time across activities: students spent 10 percent less time drafting and more time reading and searching for information. AI access also substantially increased enjoyment of the learning task, with treated students reporting 13 percent higher enjoyment ratings. Finally, AI access during the learning phase increased academic integrity violations during the test, suggesting that exposure to AI may erode compliance with rules prohibiting its use.

Our results show that off-the-shelf generative AI can improve learning. These findings provide proof of concept that AI can enhance multiple dimensions of learning, and that their magnitude depends on how AI reshapes the production function of learning. Yet this does not imply that widespread AI adoption will improve learning overall. We study one academic task—learning a new concept and demonstrating understanding through writing—common and important in undergraduate education, but only one of many academic tasks. Students use AI across many other academic activities, and possibly to save time in some of them: automating problem sets, generating first drafts, or summarizing readings. If AI primarily substitutes for cognitive effort in these settings, the net effect on learning could be negative even if the effect on any individual task is positive.

Our results contribute to a growing literature on the learning impacts of generative AI. We contextualize our effect sizes within this literature in the main text, but note two challenges in interpreting existing evidence—one on each side of the experimental design.

On the treatment side, many studies evaluate AI systems that are scaffolded, restricted, or otherwise customized beyond a standard off-the-shelf chatbot.[1] These designs are informative about the particular system studied, but less directly informative about the general-purpose chatbots that students often use in practice. On the control side, some studies compare AI against active alternatives rather than a business-as-usual counterfactual, making it difficult to isolate AI's contribution to learning.[2] Our design addresses both challenges: treated students receive off-the-shelf ChatGPT without customization, training, or safeguards, while control students learn the same material without any AI access. Beyond this design choice, we contribute by measuring knowledge persistence, by examining how AI reshapes the production function of learning—including time reallocation and engagement—and by assessing multiple learning dimensions, rather than relying on test scores alone.

We also contribute to the literature on how access to AI affects human performance. Prior studies document substantial productivity gains from AI access in professional writing, software development, consulting, and customer support, using a mix of online experiments, controlled task experiments, field experiments, and real-world workplace deployments (Noy and Zhang, 2023; Peng et al., 2023; Dell'Acqua et al., 2023; Brynjolfsson et al., 2025; Cui et al., 2026). Yet most of this literature measures short-run task performance—how well workers perform tasks using AI—rather than skill accumulation—how effectively individuals acquire durable new knowledge. The few studies that speak more directly to learning in work settings suggest mixed possibilities: Brynjolfsson et al. (2025) provide evidence that AI assistance facilitates worker learning among customer-support agents, while Budzyń et al. (2025) document patterns consistent with deskilling among endoscopists after routine AI exposure. We provide experimental evidence that AI access can improve skill accumulation in an academic context, identifying a channel through which AI may raise long-run productivity beyond its immediate effects on task performance.

Finally, we contribute to the broader literature on digital technologies and student learning. Prior research has examined laptops and personal computers (Carter et al.,

[1] For example, Bastani et al. (2025) test GPT-4 with guardrails that provide teacher-designed hints and avoid direct solution provision; Xu et al. (2025) incorporate metacognitive support into a generative-AI learning environment; Kim et al. (2025) restrict AI to a post-solution tutoring role; De Simone et al. (2025) study a structured, teacher-guided implementation of generative AI in after-school tutoring; Kumar et al. (2023) customize GPT-4 with a hidden pre-prompt; and others build bespoke tutoring systems with explicit scaffolding and prompt engineering (e.g., Kestin et al., 2025; LearnLM Team, 2025).

[2] For example, Kestin et al. (2025) compare AI against in-class active learning, LearnLM Team (2025) against static pre-written hints and human tutoring, Lira et al. (2025) against professional editor feedback, and Kreijkes et al. (2025) against note-taking.

2017; Malamud and Pop-Eleches, 2011; Cristia et al., 2017), online courses (Figlio et al., 2013; Bettinger et al., 2017), internet access (Vigdor et al., 2014; Dettling et al., 2018), and interactive classroom technologies such as clickers and whiteboards (Caldwell, 2007; Lewin et al., 2008). A central lesson from this literature is that expanding access to general-purpose technologies often yields limited, mixed, or sometimes negative effects on academic performance, whereas larger and more consistent gains tend to arise from computer-assisted learning systems that tailor content to students' learning levels, diagnose starting points, and provide immediate feedback (see Bulman and Fairlie, 2016; Escueta et al., 2020, for reviews). Generative AI extends this logic: it can engage in natural-language dialogue, generate personalized explanations, and respond flexibly to students' questions in real time. But whether these capabilities translate into learning gains depends on how students actually use the technology. Our experiment provides direct evidence on this question.

2 Context: Use of Generative AI in Higher Education

ChatGPT's release triggered the most rapid technology adoption in higher education history. Recent surveys document exceptionally rapid diffusion of AI among college students, with studies consistently reporting adoption rates exceeding 50 percent within ChatGPT's first year (Stöhr et al., 2024; Ravšelj et al., 2025). At Middlebury College, where our experiment takes place, adoption reached over 80 percent at the time we conducted the experiment (Contractor and Reyes, 2025).

These high adoption rates are partly fueled by positive student perceptions of AI's impact on learning. Across multiple studies, majorities of students report that generative AI improves their understanding of course materials, enhances their learning ability, and helps them complete assignments quicker (Stöhr et al., 2024; Ravšelj et al., 2025), a finding that also holds in our setting (Contractor and Reyes, 2025). Yet students have been shown to hold biased beliefs in educational settings (Jensen, 2010; Wiswall and Zafar, 2015), raising the possibility that widespread positive beliefs about AI may not reflect its actual effects on learning.[3]

Ex ante, predicting how AI will affect learning outcomes is difficult because generative

[3]In a workplace context, Becker et al. (2025) find that experienced open-source developers predicted AI tools would speed them up by 24 percent, and even after the study believed AI had accelerated their work by roughly 20 percent. Yet the actual measured effect was a 19 percent *slowdown*. This disconnect between perceived and actual productivity effects motivates our examination of whether students' positive beliefs about AI and learning are similarly miscalibrated.

AI is a general-purpose technology that students employ for many purposes. One useful framework (Brynjolfsson and Mitchell, 2017; Acemoglu and Restrepo, 2019) groups these uses into "augmentation" (AI works *with* the student) versus "automation" (AI does the work *for* the student. Augmentation includes AI explaining confusing concepts or giving feedback on drafts. Automation includes AI generating essays or problem set solutions. If students primarily use AI to automate cognitive tasks, it may reduce the work necessary for learning. Conversely, if students use AI to augment their understanding, it may lead to more learning. Contractor and Reyes (2025) find that students use AI for both purposes, a pattern consistent actual usage data from Claude and ChatGPT conversation logs (Handa et al., 2025; OpenAI, 2025) and from an analysis of ChatGPT interaction logs among undergraduate students at another U.S. university (Ammari et al., 2025), making the net impact on learning theoretically ambiguous.

The combination of rapid adoption, positive but potentially wrong student beliefs, and theoretically ambiguous effects motivates our experiment.

3 Experimental Design

This section describes the experimental design and implementation. The experiment consists of two in-person sessions conducted approximately one week apart. Screenshots of our surveys and experimental design are available in the Supplementary Materials.

3.1 Session One: Knowledge Production and Initial Learning

Overview. The first session employs a between-subjects design with random assignment to either an AI-allowed or AI-forbidden condition. We scheduled eight time slots with two concurrent sessions per slot, one for each treatment arm, conducted in separate computer labs. For each time slot, we randomly assigned one lab to the AI-allowed condition and the other to the AI-forbidden condition. Students who signed up for a given time slot were randomly assigned to a specific computer in one of these two labs. All computer workstations were equipped with privacy dividers to minimize distractions and prevent participants from viewing other screens (see Appendix Figure A1).

Session Structure. The first session follows a fixed 60-minute structure with the following components (Figure 1, Panel A).

Welcome and Instructions. Each session begins with a staff member reading a condition-specific welcome script. We tell participants that their primary task involves learning about a pre-specified topic and that they will demonstrate their understanding through a series of assessments. To ensure comprehension of the experimental procedures, participants must correctly answer a series of understanding checks, with laboratory staff providing additional clarification if needed. Participants answered an average of 4.77 out of 5 comprehension questions correctly, and the median participant answered all questions correctly.

Topic Assignment and Baseline Assessment. We selected three topics rich in factual content about which most undergraduate students have limited prior knowledge: blockchain technology, carbon capture systems, and CRISPR gene editing. Participants are randomly assigned to one of these three topics and cannot change their assignment.

After topic assignment, participants complete a five-question multiple-choice baseline test to assess their prior knowledge of the assigned topic before the learning phase. In all tests, the order of questions and answer options within each question is randomized. Participants are prohibited from using external resources during tests.

Learning Phase. Following the baseline test, participants enter the learning phase. During this phase, they have up to 35 minutes to learn about their assigned topic. We instruct them to "use the same learning approach you would typically follow for a college assignment" and emphasize that their performance affects their payoffs (as described below).

During this phase, participants also write an approximately 500-word analytical essay demonstrating their understanding of the topic. For example, a participant assigned to carbon capture may be asked to "analyze the three main barriers to scaling carbon capture technologies (technical limitations, economic costs, and policy challenges)," identify which barrier is most critical, and support the analysis with specific examples. More broadly, the prompts ask participants to apply critical thinking skills by analyzing relationships between concepts, comparing different elements, and supporting their analysis with specific evidence (Appendix B.2 contains complete prompt texts). For each topic, we developed two different prompts and randomly assigned each participant one prompt for Session One and the other for Session Two.

We provide all participants with a link to an introductory reading about their assigned topic. We constructed these materials to be of comparable length and reading difficulty across all three topics. Texts range from 1,253 to 1,399 words in length (Appendix Table B1). Based on an average silent reading speed of approximately 250 words per minute

for college students (Carver, 1992; Brysbaert, 2019), the estimated reading time for each text is 5.0 to 5.6 minutes. We track whether participants click on this link as a measure of engagement with provided resources. The full text of the learning materials is available in the Supplementary Materials.

Participants have complete flexibility in allocating their time across activities: reading materials, searching for information, drafting, editing, or other activities. Students can finish before the 35-minute maximum but cannot exceed it. The interface automatically advances at 35 minutes.

Post-Learning Survey. After the learning phase, participants complete a brief survey measuring self-assessed current knowledge of the topic (0-10 scale), time allocation across different activities during the learning phase, subjective task assessment (enjoyment, feeling skilled or effective), whether they had previously done similar tasks in their current academic year, and an attention check.

Post-Learning Test. The session concludes with five multiple-choice questions testing factual knowledge and conceptual understanding of the assigned topic. These questions differ from the baseline test. Participants complete this assessment without access to any external resources, internet, or AI tools regardless of their treatment condition.

Waiting Room. Participants who complete all tasks before the 60-minute session duration are directed to a virtual waiting room. The Finish Session 1 button is enabled only after 50 minutes have elapsed from the session start time. Participants are informed that their responses will not be recorded until they click this button. This prevents early departures from disrupting others or creating social pressure effects. During this waiting period, participants may browse the internet freely.

Treatment. The experimental manipulation centers on access to generative AI tools during the learning phase. The treatment has two components: explicit instructions and platform access.

Participants in the *AI-allowed* condition receive explicit instructions permitting use of any generative AI tools during the learning phase. These instructions are embedded in the welcome script that proctors read aloud at the start of each session. The script states: "[I]f you typically use generative AI tools such as ChatGPT, feel free to use them here as well. Using ChatGPT or other generative AI is completely allowed. We provide you with a ChatGPT account, already logged in one of the tabs, so you don't need to use your personal account." The instructions are also communicated in written form through the

survey interface (Appendix Figure A2, Panel A). To facilitate usage, each computer has three browser tabs already open: a dedicated ChatGPT (GPT-4o) account, Wikipedia, and the Middlebury College Library website.

Participants in the *AI-forbidden* condition are instructed: "You are not allowed to use generative AI tools such as ChatGPT, Claude, or other AI assistants." These instructions are part of the welcome script and communicated in written form through the survey interface (Appendix Figure A2, Panel B). Computers assigned to AI-forbidden participants have three browser tabs already open: Google, Wikipedia, and the Middlebury College Library website.

Compliance Monitoring. We enforce treatment compliance through multiple mechanisms. First, two proctors per lab directly observe participants throughout the session using classroom seating charts to record any unauthorized resource use (Appendix Figure A3). Proctors note instances of students accessing AI tools in the control condition or violating other experimental protocols such as accessing external resources during the tests. Second, the exit survey elicits self-reported AI usage during the learning phase, including which specific AI tools were used (ChatGPT, Claude, Gemini, etc.). Third, we periodically capture participants' writing interfaces to detect potential large pasting events—sudden appearances of large text blocks that may indicate copying from external sources (Noy and Zhang, 2023).

Performance Incentives. To incentivize effort, we implement a lottery-based compensation structure. Participants earn lottery tickets based on performance. They receive one ticket for each correct response on each test. They also receive one lottery ticket for each point on their rounded quality score. At the conclusion of the experiment, we conduct a drawing in which 30 lottery tickets are randomly selected, with each winning ticket worth $100. Participants can win multiple prizes if multiple of their tickets are drawn.

3.2 Session Two: Knowledge Retention

Session Two takes place approximately one week after Session One.[4] This session measures knowledge retention when all participants—regardless of Session One treatment

[4]When scheduling, we encouraged participants to sign up for a Session Two slot approximately one week after their Session One slot, but allowed flexibility to accommodate scheduling constraints. The average gap between sessions was 7.0 days. Over 61 percent of students scheduled their second session exactly 7 days after Session One, and over 80 percent attended within 6 to 8 days.

assignment—work without access to AI or external resources.

Session Structure. The session includes two assessments presented in randomized order (Figure 1, Panel B): a 10-item multiple-choice knowledge test with questions distinct from Session One, and a 20-minute analytical essay using the complementary prompt developed for each topic.

Session Two concludes with a questionnaire measuring beliefs about AI's impact on learning, understanding, course grades, and time spent on academics; self-reported AI skills (Likert scale) and frequency of AI use for different academic tasks (never to daily); when participants started using AI and specific academic applications (brainstorming, writing, editing, research, coding, etc.); retrospective Session One behavior including self-reported use of AI during Session One, which tools were used, and for what purposes (explaining concepts, summarizing readings, drafting, proofreading, editing); whether participants studied the topic or accessed materials between sessions; and an open-ended question about the impact of generative AI on learning.[5]

4 Data and Research Design

4.1 Sample and Recruitment

The experiment took place at Middlebury College during Spring 2025. We recruited undergraduate participants through campus-wide email announcements, posted flyers, student group chats, and a dedicated recruitment website. To minimize selection concerns, we framed the study broadly as "a study on learning" without mentioning AI in any recruitment materials.

Students were required to attend two in-person sessions approximately one week apart. We implemented several design features to minimize attrition. First, we structured com-

[5]We note four main deviations from the experimental protocol that occurred during early time slots. First, the initial Session One verbal welcome script did not mention AI permissions or restrictions; participants learned their treatment condition only through written platform instructions. We introduced verbal acknowledgment of treatment conditions starting with the second Session One time slot. Second, the initial Session One included a timer that malfunctioned, displaying available time after the 35-minute limit had elapsed. We removed the timer for subsequent sessions. Third, the first Session Two time slot lacked a waiting room, requiring participants to manually track time before leaving. We introduced a virtual waiting room for all following sessions. Fourth, a technical glitch during one Session Two time slot prevented participants from completing the essay on the platform. We added extra time and asked affected participants to write their essays in Microsoft Word, manually incorporating these responses into the dataset. All analyses include session fixed effects to account for these protocol variations.

pensation to incentivize completion of both sessions: participants receive $5 for completing Session One alone, but $50 for completing both sessions, with all payments disbursed at the conclusion of Session Two. Second, participants received email and text reminders on the day of their scheduled session. Third, we provided flexibility in the timing of the second session rather than requiring exactly seven days between sessions. We also allowed students to reschedule if conflicts arose.

4.2 Summary Statistics and Balance

Table 1 reports summary statistics for three groups: the 256 students who completed the intake survey (column 1), the 210 (82.0 percent) who attended Session One (column 2), and the 204 (79.7 percent) who attended both sessions (column 3). Among Session One attendees, 97.1 percent returned for Session Two. We focus on column 2. The typical student is 20.1 years old; 35.2 percent are male, 49.5 percent white, 26.7 percent international, and just over half (53.8 percent) attended a public high school (Panel A). The sample spans all four undergraduate years, with 37.6 percent first-years, and represents diverse academic backgrounds: 31.0 percent study natural sciences, 35.2 percent social sciences (Panel B). Academic performance is strong, with a mean GPA of 3.68 and an average SAT score of 1386, approximately the 93rd percentile of test takers (College Board, 2025).[6] Students self-report little baseline knowledge of their assigned topic (1.5 on a 0–10 scale) and, consistent with this, answer only 30.4 percent of baseline multiple-choice questions correctly.

The sample is balanced on most observable characteristics across treatment groups. Table 2 compares mean characteristics of students in the AI-forbidden versus AI-allowed conditions (columns 1 and 2). Only the coefficient on age is statistically significant at the ten percent level (column 3). An F-test of joint balance cannot reject the null that all differences equal zero at conventional levels. Most importantly, measures of academic ability—strong predictors of learning outcomes—are well-balanced: mean GPA (3.71 versus 3.66) and SAT scores (1391 versus 1380) are nearly identical across groups. To address any residual imbalance, we use the double-lasso procedure (Belloni et al., 2014), as described below. We find no differential attendance in Session One (82.5 versus 82.2 percent) and no differential attrition between sessions.

[6]We asked participants for both SAT and ACT scores. Most students took the SAT. When students provided ACT but not SAT scores, we use concordance tables to convert ACT scores to SAT equivalents.

4.3 Regression Model

We estimate linear models of the form:

$$Y_i = \alpha + \beta \text{AI-Allowed}_i + \mathbf{X}_i'\gamma + \varepsilon_i, \tag{1}$$

where Y_i is an outcome for participant i, AI-Allowed_i is an indicator equal to one if the participant was randomly assigned to the AI-allowed condition, $\mathbf{X}_i$ is a vector of baseline control variables to improve precision, and ε_i is an error term. The vector $\mathbf{X}_i$ contains the randomization dummies as well as additional controls selected using the double-lasso procedure (Belloni et al., 2014). This procedure selects controls that predict either the outcome or the treatment assignment from a pool of potential covariates.[7] We report heteroskedasticity-robust standard errors.

The primary coefficient of interest, β, measures the causal effect of AI access on various outcomes. This intent-to-treat (ITT) estimate captures the effect of being *allowed* to use AI during the learning phase, regardless of whether students actually used the AI tools. This may be the relevant parameter for policy evaluation, as institutions can control whether students are permitted to use AI but cannot enforce actual usage. Still, to learn how AI usage affects the learning of users, we complement our ITT estimates with a treatment-on-treated (TOT) analysis that measures the effect of actual AI usage rather than just assignment. We instrument actual AI use with the randomly assigned treatment status (AI-allowed). The resulting two-stage least squares (2SLS) estimates provide the local average treatment effect (LATE) for participants whose AI usage was influenced by their treatment assignment.

4.4 Main Outcomes

We estimate equation (1) separately for different categories of outcomes:

[7] The pool of potential controls includes demographic characteristics (age, gender, race/ethnicity indicators for Black, Latino, Asian, international student status, public versus private high school attendance), academic background (cohort fixed effects, declared major status, field of study indicators for natural sciences, social sciences, and humanities/arts, self-reported study hours per week, college GPA, SAT score, whether an admissions test was taken, plus quintile bins of SAT scores, college GPA, and high school GPA to allow for flexible functional forms; for the 24 percent of students who did not take a standardized admissions test, we impute SAT scores from baseline covariates to assign them to bins, as described in Appendix B.5), baseline measures (pre-learning self-assessed knowledge, pre-learning test score, prior experience with similar tasks, comprehension check score), and experimental design features (topic fixed effects, writing prompt version, and lab location).

AI Usage. As a measure of a "first-stage", we analyze whether students used generative AI during the learning phase. Our primary measure is an indicator equal to one if at least one prompt was found in the ChatGPT account provided to the student during Session One or if a proctor directly observed the student using ChatGPT. We also measure the intensity of AI use through the number of prompts sent to the provided ChatGPT account and the number of distinct conversations (i.e., separate chat threads) initiated during the session. Self-reported measures capture whether students used any generative AI tool, whether they specifically used ChatGPT, and whether they used alternative AI models (Claude, Gemini, Copilot, DeepSeek, or Llama).

Test-Based Learning Outcomes. We measure knowledge acquisition and retention using both self-reported and objective measures. Self-assessed knowledge captures students' subjective evaluation of their understanding on a 0-10 scale, where 0 indicates "I know nothing about this topic" and 10 indicates "I am an expert."[8] The fraction correct is the proportion of multiple-choice questions answered correctly (five questions in Session One, ten in Session Two). We construct standardized test scores by normalizing raw scores to have mean zero and standard deviation (SD) one in the control group. We also define binary indicators for scoring above various performance thresholds to examine treatment effects across the score distribution.

Essay Evaluation. Writing assessments capture higher-order skills such as analytical reasoning, synthesis, and argumentation. Each essay is independently evaluated by three human graders recruited through Prolific, blind to treatment condition and participant identity, on a 0–10 scale. Graders provide a holistic overall quality rating and scores on five specific dimensions: factual accuracy, use of evidence and examples, relevance to the prompt, structure and organization, and writing style and clarity (see Appendix B.3 for the grading instrument). We also construct a quality index by standardizing each dimension to have mean zero and standard deviation one in the control group and averaging across the five dimensions. We complement human grading with AI grading using Claude Opus, which evaluates each essay on the same dimensions using the same rubric.[9] For human-graded

[8] This measure may capture dimensions of learning that a limited set of multiple-choice questions cannot test, though we acknowledge that it requires accurate self-knowledge and is subject to the usual biases of subjective questions, such as social desirability bias.

[9] Our use of AI grading is motivated by a growing literature validating LLMs as evaluators of text quality (Chiang and Lee, 2023; Liu et al., 2023; Zheng et al., 2023), and by the fact that our learning topics (blockchain, carbon capture, CRISPR) are technical subjects on which human graders recruited

outcomes, we use the individual grader-essay pair as the unit of observation, including grader and batch fixed effects to absorb grader leniency differences; AI-graded outcomes are at the student level.

Writing Style. In addition to quality ratings, we analyze three dimensions of participants' writing style: length, readability, and lexical diversity. We measure length through token, word, and sentence counts. Readability is captured by sentence length, syllables per word, the Flesch-Kincaid grade level, and the Flesch Reading Ease score. Lexical diversity is measured through the type-token ratio (ratio of unique words to total words) and the hapax proportion (share of words appearing only once). To reduce dimensionality, we standardize each component to have mean zero and standard deviation one in the control group and aggregate them into three indices: a length index, a readability index (with difficulty measures reversed so that higher values indicate easier text), and a lexical diversity index. We also create an indicator for empty essays.

To examine whether AI access homogenizes student writing, following Brynjolfsson et al. (2025), we compute sentence embeddings using the `all-mpnet-base-v2` transformer model. We measure within-group textual similarity as the average pairwise cosine similarity between a student's essay and all other essays in the same treatment group, topic, and prompt cell, and source text similarity as the cosine similarity between the essay and the provided reading material.[10] Finally, we use Pangram, a commercial AI content detector (Emi, 2024; Masrour et al., 2025; Thai et al., 2026), to measure the fraction of text classified as AI-generated and the fraction flagged as plagiarized.[11]

5 Immediate Effects of AI Access on Learning and Writing

5.1 First Stage: AI Adoption and Usage Patterns

We begin by examining AI access's impact on AI usage during the learning phase. Table 3 reports estimates from equation (1) using several measures of AI usage. The first row

from a general population may have limited expertise. The correlation between average human and AI scores for overall essay quality is 0.48 in Session One and 0.57 in Session Two (both $p < 0.001$). See Appendix B.4 for the AI grading procedure and prompts.

[10] As a validation exercise, we confirm that the embeddings capture meaningful content variation: average within-topic cosine similarity is 0.75, compared to 0.15 across topics, and essays responding to the same prompt are more similar to each other (0.79) than essays on the same topic but a different prompt (0.71).

[11] Jabarian and Imas (2025) evaluate four AI text detectors and find that Pangram achieves near-zero false positive and false negative rates, even when AI-generated text is modified using "humanizer" tools.

presents the revealed preference measure based on activity in the ChatGPT accounts provided to treatment group participants. The remaining rows present self-reported measures. Figure 2 illustrates these first-stage effects.

The experimental manipulation generated substantial variation in AI usage. Assignment to the AI-allowed group increased AI usage by 67.9 percentage points (pp) as measured by activity in the provided ChatGPT account ($p < 0.01$). This effect represents a large increase from the near-zero baseline in the control group. Self-reported measures yield similar though slightly attenuated results: overall generative AI usage increased by 59.9 pp ($p < 0.01$). Virtually all adoption occurred through ChatGPT—usage increased by 60.2 pp ($p < 0.01$), while usage of other AI models remained negligible at 0.9 pp.[12]

We examine which student characteristics predict AI take-up among those assigned to the treatment group (Appendix Table A1). The strongest predictors are prior AI experience and self-assessed proficiency: students who report more frequent AI use during the academic semester and higher AI proficiency are more likely to use the provided ChatGPT account. Students who perceive larger benefits from AI are also more likely to adopt. More suggestively, non-white and low-GPA students show higher take-up rates than white and higher-achieving students.

Having established that AI access substantially increased usage, we examine how treated students used these tools. Figure 3 reports the fraction of treated students who used AI for each of six purposes, measured through self-reports (Panel A) and through classification of actual ChatGPT conversation logs (Panel B). Both sources agree that explaining concepts was the most common use (44 and 57 percent, respectively), followed by summarizing readings and drafting responses. The revealed data shows higher rates of drafting, editing, and "other" uses than self-reports, suggesting that students underreport uses related to text generation.[13] Over 89 percent of students in the AI-allowed group reported that AI was somewhat or very helpful for the learning task.

5.2 Effects on Learning

We next examine the impact of AI access on immediate learning outcomes measured at the end of Session One. AI access produced significant improvements in actual learning but

[12]This concentration on ChatGPT is consistent with surveys of adoption among college students (Contractor and Reyes, 2025) and the U.S. working-age population (Bick et al., 2026).

[13]The high rate of "other" uses in the revealed data largely reflects conversational turns (e.g., "yes," "sure"), word count requests, synonym lookups, and off-topic questions—prompts that students would not consider a distinct use of AI when responding to a survey.

not in perceived learning (Table 4). Self-assessed knowledge decreased by -0.03 points, an effect that is not statistically significant. Performance on the objective knowledge test improved substantially. Students in the AI-allowed condition answered 6.3 pp more questions correctly (column 2, $p < 0.05$), equivalent to an 11 percent improvement over the control mean of 56.3 percent correct. When standardized, this translates to a 0.25 SD improvement in test scores (column 2, $p < 0.05$). The disconnect between subjective and objective measures suggests that students may not fully perceive the learning gains facilitated by AI access. These effects are similar across all three topics (Appendix Table A4) and robust to excluding students who failed attention or comprehension checks, or who studied the topic between sessions (Appendix Table A5).

The learning gains from AI access are concentrated among middle-performing students. Table 5 reports treatment effects on the probability of scoring above various performance thresholds. Students with AI access were 7.9 pp more likely to score at least 40 percent correct and 11.5 pp more likely to score at least 60 percent correct ($p < 0.10$). AI had minimal effects on reaching the 80 percent threshold or achieving perfect scores. These effects are illustrated in Figure 4, which shows that AI access shifts the middle of the performance distribution rightward while leaving the tails largely unchanged. This pattern is consistent with previous work showing that AI disproportionately benefits workers in the lower half of the performance distribution (Noy and Zhang, 2023; Brynjolfsson et al., 2025).

Figure 6 examines whether these distributional effects reflect differences in baseline student ability, plotting treatment effects on test scores by GPA and SAT quartile. The clearest heterogeneity emerges along academic preparedness, though the two proxies tell different stories. Low-GPA students experience the largest immediate test score gains (0.40 SD, $p < 0.05$), but these effects fade by Session Two (0.03 SD). High-GPA students show the opposite trajectory, with modest Session One effects that grow by Session Two. The SAT split yields a similar pattern: high-SAT students benefit across nearly all outcomes, with effects that strengthen over time, while low-SAT students show positive but imprecise Session One effects and near-zero Session Two effects (Appendix Table A6 reports the full set of interaction effects). AI access thus narrows achievement gaps in the short run—lower-performing students gain the most from immediate AI assistance—but may widen them over time, as academically stronger students appear better equipped to retain what they learn with AI assistance.

Figure 5 places our estimates in the experimental literature on AI and learning. We

surveyed 20 papers and applied strict inclusion criteria—randomized designs, unassisted assessment outcomes, no-AI controls, and reportable effect sizes—to select four studies with nine comparable estimates (Appendix B.7). Our Session One effect of 0.25 SD falls near the median of the curated distribution, which ranges from −0.19 SD (Bastani et al., 2025) to 0.42 SD (Lehmann et al., 2025). Studies providing unrestricted chatbot access—the design closest to ours—yield moderate and occasionally negative effects (Bastani et al., 2025), while structured coaching interventions produce the largest gains (Lira et al., 2025). Our retention effect of 0.27 SD falls within the same range, though few studies measure delayed outcomes. Studies finding negative effects after initial gains during AI-assisted practice (Bastani et al., 2025) typically remove AI access between sessions, as we do, consistent with the persistence of learning gains depending on how deeply students engage with AI-generated content.

5.3 Effects on Writing Quality

We next examine whether AI access affected the substantive quality of students' essays, as assessed by independent graders blind to treatment condition. Because Session One essays were written during the learning phase—when treated students had AI access—these estimates reflect both direct AI assistance and any changes in how students approach the writing task; we isolate the learning component in Section 6, when all students write without AI access.

Table 6 reports treatment effects on each of the six grading dimensions, standardized to have mean zero and standard deviation one in the control group. Figure 7 presents ITT effects visually in standard deviation units.

AI access had no measurable impact on essay quality across all six dimensions, with the point estimates being small in magnitude and not statistically significant. The largest improvements appear in the presentation dimensions: writing style and clarity, and structure and organization. Substantive dimensions—factual accuracy, use of evidence, and relevance to the prompt—also show positive point estimates. Overall quality, the grader's holistic assessment, shows a similar pattern. Appendix Table A3 examines distributional effects on overall essay quality.

5.4 Effects on Writing Style

We also examine how AI access affected participants' writing style during Session One. As with essay quality, these effects combine direct AI assistance with changes in student

behavior; Section 6 examines persistence. Table 7 presents ITT and TOT estimates for five writing indices: length, readability, lexical diversity, within-group textual similarity, and source text similarity. Figure 7 presents visual evidence on these treatment effects in standard deviation units.

AI access suggestively led students to produce longer essays written in a clearer, more accessible style—using simpler words and shorter sentences. The length index increased by 0.15 SD (column 2), corresponding to about 15 additional tokens, or approximately a 7 percent increase from the control mean of 203 tokens (Appendix Table A2). The readability index increased by 0.09 SD (column 2), reflecting improvements across multiple dimensions: the Flesch Reading Ease score increased by 3.9 points and mean sentence length fell by 4.6 words (Appendix Table A2). The lexical diversity index shows a small positive effect of 0.08 SD (column 2). While these point estimates are imprecisely estimated and do not reach statistical significance, they suggest a pattern of AI-facilitated changes in writing style toward greater clarity and accessibility.

We find no evidence that AI access increases textual homogeneity. Within-group cosine similarity shows a small, negative, and statistically insignificant effect (column 2). If anything, AI-allowed essays are slightly *less* similar to one another than AI-forbidden essays, though the differences are not distinguishable from zero. This contrasts with the convergence patterns documented by Brynjolfsson et al. (2025) among customer support agents, possibly reflecting differences between a workplace setting with repeated interactions and a one-shot academic writing task.

Finally, we examine whether AI access left detectable traces in students' writing using Pangram, a commercial AI content detector. The fraction of text classified as AI-generated increased by 9.4 pp in the AI-allowed group ($p < 0.05$), an 73 percent increase over the control mean of 12.8 percent. The control mean itself is non-trivial: 12.8 percent of control group text is flagged as AI-generated despite these students having no access to AI tools. This baseline rate reflects either false positives or students whose natural writing style resembles AI-generated text—a pattern consistent with recent evidence that exposure to large language models has begun shifting human communication patterns toward AI-like prose (Yakura et al., 2025). In Session Two, when no students had AI access, the treatment effect drops to zero, and the control mean falls to 4.1 percent, suggesting that the Session One flags in the control group partly reflected topic-specific or prompt-specific stylistic patterns rather than stable individual writing traits.

5.5 Heterogeneity of Treatment Effects

We next examine how treatment effects vary by the type of AI interaction and by student characteristics. Our experiment is not powered to detect subgroup differences, so these patterns should be interpreted as suggestive.

Automation versus Augmentation. Table 8 distinguishes between automation and augmentation users, classified from ChatGPT conversation logs (see Appendix B.6 and Appendix Figure A6). The two types of AI interaction produce sharply different effects. Augmentation users—those who primarily sought explanations or feedback—show significant test score gains (0.32 SD, $p < 0.05$) but no improvement in essay quality, consistent with deeper conceptual engagement translating into better test performance. Automation users—those who primarily used AI to draft text—show the opposite pattern: large Session One quality gains (0.69 SD for overall quality and 0.60 SD for the quality index, both $p < 0.05$) that vanish entirely by Session Two, consistent with AI-produced text inflating quality measures without durable learning. These patterns reinforce the distinction between augmentation and automation as a key moderator of AI's educational effects.

AI Experience and Demographics. Prior AI experience is also a strong moderator (Appendix Table A6). Frequent AI users show large and persistent test score gains (0.50 SD in Session One and 0.77 SD in Session Two, both $p < 0.01$), while infrequent users show no significant test score effects in either session. The pattern for essay quality differs. Early adopters and high-proficiency students show significant Session One quality gains (0.67 SD and 0.58 SD for overall quality, both $p < 0.05$), consistent with these students using AI effectively for writing. Late adopters and low-proficiency students show no Session One quality effects but significant Session Two gains (0.49 SD and 0.27 SD for overall quality). This catch-up pattern may reflect delayed learning: students less familiar with AI initially struggle to use it productively but absorb enough from the experience to improve their writing one week later.

Treatment effects are broadly similar across demographic groups. We find no robust differences by gender, race, nationality, or field of study. Some suggestive patterns emerge—female students show larger Session One test score gains (0.36 SD, $p < 0.05$) than male students (0.07 SD), and international students show larger Session Two quality gains than domestic students—but none of these differences are statistically distinguishable given our sample size.

5.6 Mechanisms: How AI Transforms the Learning Process

We examine four potential mechanisms through which AI access affected learning: engagement with the learning task, time reallocation across different learning activities, learning experience, and academic integrity norms.

Engagement. Panel A of Table 9 examines whether AI access altered engagement with the learning task. We measure engagement through four outcomes: whether students opened the provided reading materials, time spent on the learning task (both objectively recorded via Qualtrics and self-reported), and whether participants correctly remembered which of the three topics they were assigned in Session One.

AI access had no detectable effects on engagement metrics. Access to the reading material did not differ between groups. Time measures showed no meaningful differences: objectively recorded time decreased by 0.2 minutes, while self-reported time decreased by 0.6 minutes, though neither effect is statistically significant. Topic recall was nearly universal and identical across groups: 96 percent of control group students correctly remembered their assigned topic, with a zero treatment effect.

Time Allocation across Learning Activities. While total learning time remained unchanged, students may have reallocated their time across different activities. Figure 8 plots the treatment effects on time spent across seven learning activities: drafting, editing, note-taking, organizing, searching for information, reading, and other activities.

AI access appears to have reallocated time across learning activities, with a suggestive pattern of substitution away from writing and toward information-seeking. Time spent drafting decreased by 1.08 minutes, representing a 10 percent reduction from the control mean of 11.07 minutes. These reductions in writing activities were partially offset by increases in information-seeking: reading time increased by 0.71 minutes, while time spent searching increased by 0.38 minutes, though these effects are imprecisely estimated and not statistically significant at the conventional levels. Other activities showed minimal changes. This reallocation is consistent with a broader pattern documented in professional settings, where AI shifts effort from task execution to task stewardship (Lee et al., 2025). Mozannar et al. (2024) find that software developers using GitHub Copilot spend more than half their coding time verifying and editing AI-generated suggestions rather than writing code, suggesting that AI access redirects effort from production to evaluation.

Learning Experience. Panel B of Table 9 examines whether AI access affected participants' subjective experience of the learning task, measuring task enjoyment and perceived skillfulness on both continuous (0–10) and binary (above-median) scales.

AI access substantially increased participants' enjoyment of the learning task. Task enjoyment increased by 0.68 points on the 10-point scale ($p < 0.10$), representing a 13 percent improvement from the control mean of 5.22. This effect translates to a 15.2 pp increase in the likelihood of reporting above-median enjoyment ($p < 0.05$). In contrast, AI access had minimal effects on how skilled or effective students felt during the task. Perceived effectiveness increased by only 0.04 points on the continuous scale and 4.5 pp on the binary measure, with neither effect approaching statistical significance. These findings align with Noy and Zhang (2023), who find that ChatGPT access increases task satisfaction by 0.40 SD and self-efficacy by 0.20 SD among professional writers, though this last effect is not statistically significant. We find a similar-sized effect on task enjoyment (0.30 SD) and also a positive but not significant effect on self-efficacy (0.05 SD).

Academic Integrity. Panel C of Table 9 investigates how AI access influenced academic integrity during the knowledge tests, where all participants were prohibited from using external resources. We examine three measures of unauthorized resource use: whether proctors caught students cheating, self-reported unauthorized resource use, and a combined indicator for either measure detecting cheating.

AI access during the learning phase significantly increased the likelihood of rule violations during subsequent assessments. Students in the AI-allowed condition were 5.7 pp more likely to be caught by proctors accessing unauthorized resources during tests ($p < 0.05$), more than doubling the control group rate of 2.9 percent. Self-reported violations increased by 8.6 pp ($p < 0.05$), more than quadrupling the control mean of 2.0 percent. When combining both detection methods, treatment group students were 12.6 pp more likely to cheat ($p < 0.01$), representing a large increase from the 5.0 percent baseline.[14]

[14]This finding is consistent with student perceptions documented in prior surveys. Ravšelj et al. (2025) report that a substantial fraction of students believe using ChatGPT might encourage cheating (44.9 percent), unethical behavior (32.8 percent), or plagiarism (43.5 percent), with 56.9 percent endorsing at least one of these concerns. This may be related to widespread perceptions that AI use itself constitutes cheating. For instance, Nam (2023) find that 54 percent of U.S. students view AI use on assignments as cheating or plagiarism.

Mediation Analysis. We employ a product-of-coefficients approach to quantify each mechanism's contribution to the total treatment effect on test scores (Imai et al., 2010). For each mediator M_j, we estimate the indirect effect as $\hat{\alpha}_j \times \hat{\theta}_j$, where $\hat{\alpha}_j$ is the effect of treatment on the mediator and $\hat{\theta}_j$ is the effect of the mediator on the outcome controlling for treatment. We aggregate indirect effects across all mediators within each mechanism group and obtain standard errors via a nonparametric bootstrap with 500 replications, clustering at the student level.

The total treatment effect on Session One test scores is 0.25 SD. The largest indirect pathway operates through academic integrity: increased cheating during the test accounts for 0.10 SD, or 40 percent of the total effect ($p < 0.10$), consistent with the large treatment effects on rule violations documented above. Time reallocation contributes 0.06 SD (22 percent), though imprecisely estimated, suggesting that the shift toward information-seeking activities may have facilitated learning. Learning experience makes a positive but small contribution (5 percent), consistent with AI increasing enjoyment without directly translating into deeper learning. Engagement contributes negatively (−12 percent), reflecting the small and mixed treatment effects on this margin. The remaining 45 percent of the total effect operates through a direct (unexplained) pathway from AI access to test performance, suggesting that the mechanisms through which AI improves learning are not fully captured by our mediators—such as the quality of conceptual understanding acquired through AI-assisted study or improved retention of information encountered during AI interactions.

6 Longer-Term Impacts of Access to Generative AI

6.1 Impact on Knowledge Retention

We now examine the effects of AI access during initial learning on knowledge retention measured approximately one week later in Session Two.[15] Table 4 reports estimates using the same outcomes as Session One.

The knowledge advantage from AI access persisted one week later, though with some attenuation. Self-assessed knowledge increased by 0.19 points (column 5), an effect that remains statistically insignificant. Objective test performance remained higher for students

[15]Students were not informed that they would complete additional assessments in Session Two. We find no evidence that students prepared for Session Two between sessions: only 5 percent of students reported looking up extra information about the topic, and fewer than 1 percent reported studying the topic, with no significant differences between treatment and control groups ($p = 0.18$ and $p = 0.32$, respectively).

who had AI access during initial learning. They answered 3.7 pp more questions correctly (column 5, $p < 0.10$), representing a 7.5 percent improvement over the control mean of 49.3 percent. When standardized, this translates to a 0.20 SD advantage (column 5, $p < 0.10$). Comparing these effects to Session One reveals that the learning advantage faded by approximately one-fifth: the initial 6.4 pp gap narrowed to 3.7 pp, while the standardized effect decreased from 0.25 to 0.20 SD. The persistence of a meaningful advantage one week later suggests that AI-assisted learning produces genuine, though partially decaying, knowledge gains.

We examine heterogeneity across the performance distribution in Table 5, which reports treatment effects on reaching various performance thresholds. AI access primarily helped students achieve low and medium performance levels. Treatment group students were 4.0 pp more likely to score at least 20 percent correct. The point estimate suggests a 7.7 pp increase in scoring at least 60 percent correct, though this effect is not statistically significant. Effects on reaching 40 percent correct, 80 percent correct, or achieving perfect scores are small and statistically insignificant. Figure 4 illustrates these patterns, showing that AI access continues to shift the lower portion of the distribution rightward one week later, though the effects are attenuated compared to Session One.

6.2 Persistence of Writing Quality and Style Changes

Session Two essays provide a cleaner test of whether AI exposure produced durable changes in students' writing, since all participants wrote without AI access. Table 6 (columns 4–6) reports treatment effects on the six grading dimensions, and Figure 7 presents ITT effects visually.

In contrast to Session One, where quality effects were positive but not statistically significant, Session Two reveals a clearer pattern of improvement. Human-graded overall quality increased by 0.24 points on the 0–10 scale (column 5), and AI-graded overall quality increased by 0.26 points (column 5), though neither reaches statistical significance individually. Among individual dimensions, two show statistically significant gains: use of evidence and examples improved by 0.30 points ($p < 0.10$), and writing style and clarity improved by 0.30 points ($p < 0.10$). Organization shows a similar-sized but imprecise effect (0.24 points). These results suggest that exposure to AI-generated text during Session One may have improved students' own writing in dimensions related to presentation and evidence use, even after AI access was removed. Distributional effects reinforce this interpretation: AI access increased the probability of scoring 9 or above on overall quality by 5.5

pp ($p < 0.05$), relative to a control mean of 7.5 percent (Appendix Table A3), consistent with a learning channel rather than mere AI-assisted production.

Writing style changes partially persisted as well, though these effects are imprecisely estimated (Table 7, columns 4–6). The length index increased by 0.19 SD, corresponding to about 12.8 additional tokens (Appendix Table A2), and the readability index showed modest persistence at 0.05 SD: essays exhibited easier reading scores (3.7 point increase in Flesch Reading Ease) and shorter sentences (1.8 fewer words per sentence). Lexical diversity decreased by 0.19 SD ($p < 0.10$): the Type-Token Ratio fell by 0.014 points and the hapax proportion decreased by 0.018 points. Most of these point estimates suggest some durability, but they are not statistically significant at conventional levels.

6.3 Do Students Understand AI's Impact on Their Learning?

Students' beliefs about AI's learning benefits may shape adoption decisions. In a companion survey of the same student population, Contractor and Reyes (2025) find that students overwhelmingly believe AI improves their learning, and that these beliefs are among the strongest predictors of adoption. Our experimental design allows us to test whether such beliefs are accurate by comparing perceived and actual treatment effects directly.

After completing the Session Two retention test, participants estimate: (1) how many questions they answered correctly; (2) how many they would have answered under the opposite treatment condition; (3) how many questions other students in their group answered correctly; and (4) the same for students in the opposite group. Figure 9, Panel A compares these beliefs with observed treatment effects; Panel B plots perceived against actual effects across subgroups.

Panel A reveals two facts about individual beliefs. First, students substantially overestimate their own performance: the average student answered 4.9 questions correctly, but control students believed they answered 6.3 and treated students believed 6.7. Second, the degree of overestimation depends sharply on experience with AI. AI-forbidden students predicted that AI would increase their own performance by 2.5 questions—more than six times the actual effect of 0.4 questions—and predicted a 1.4 question increase for others, more than three times the actual effect. Students who used AI developed far more calibrated beliefs: they predicted a 0.3 question benefit for themselves, closely matching the observed 0.4 question effect. Their predictions about others (1.1 questions) remain inflated but are substantially more accurate than AI-forbidden students' beliefs.

Panel B shows that this overestimation pattern holds across subgroups. For each sub-

group with at least 40 observations, we plot mean perceived against actual treatment effects. The two are positively correlated ($\rho = 0.32$), but perceived effects are systematically inflated: most subgroups report perceived gains exceeding estimated treatment effects. If students overestimate AI's benefits, they may over-rely on it in settings where it is less effective, or resist restrictions they perceive as costly.

7 Conclusion

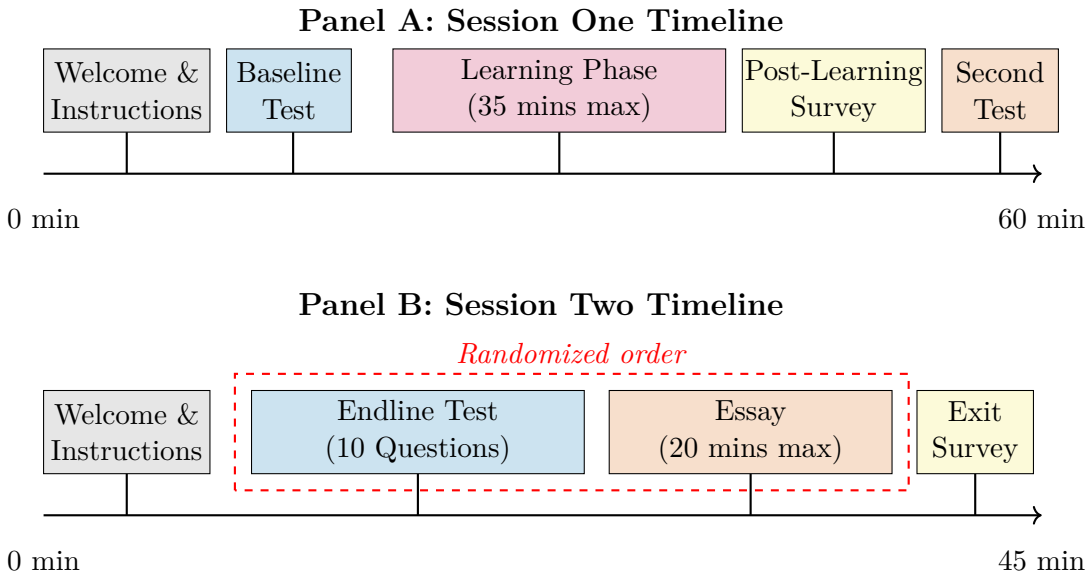
This paper provides causal evidence on how access to generative AI affects student learning. Using a randomized controlled trial with undergraduate students, we find that AI access produces significant improvements in immediate learning outcomes, with most of these gains persisting one week later. Students primarily used AI for explaining concepts rather than automating tasks, and learning gains were concentrated among middle performers rather than top students. AI access also increased rule-breaking behavior during subsequent assessments, suggesting spillover effects on academic integrity norms.

Our findings contribute to evidence that AI's educational impact depends on implementation details and usage patterns. In workplace settings, Brynjolfsson et al. (2025) find that AI helps novice call-center workers learn more effectively, while Budzyń et al. (2025) show that AI assistance during colonoscopies leads to skill deterioration. In educational settings, Bastani et al. (2025) find that unrestricted AI access harms learning in high school mathematics, but blocking solution-seeking while permitting explanation-seeking mitigates negative effects. Lehmann et al. (2025) document that students who copy-paste AI code learn less, but those requesting explanations learn more. Our results align with this pattern: AI enhances learning when students use it for conceptual understanding rather than automation.

These findings have implications for educational policy. Blanket policies prohibiting or permitting AI may be suboptimal. Our results suggest that approaches guiding students toward productive uses (concept explanation, feedback) while discouraging complete automation of cognitive tasks may be more effective. AI may help reduce achievement gaps by assisting struggling students in reaching competency thresholds. The disconnect between objective learning gains and subjective enjoyment creates potential misalignment between student preferences and educational effectiveness, particularly for institutions relying on satisfaction metrics.

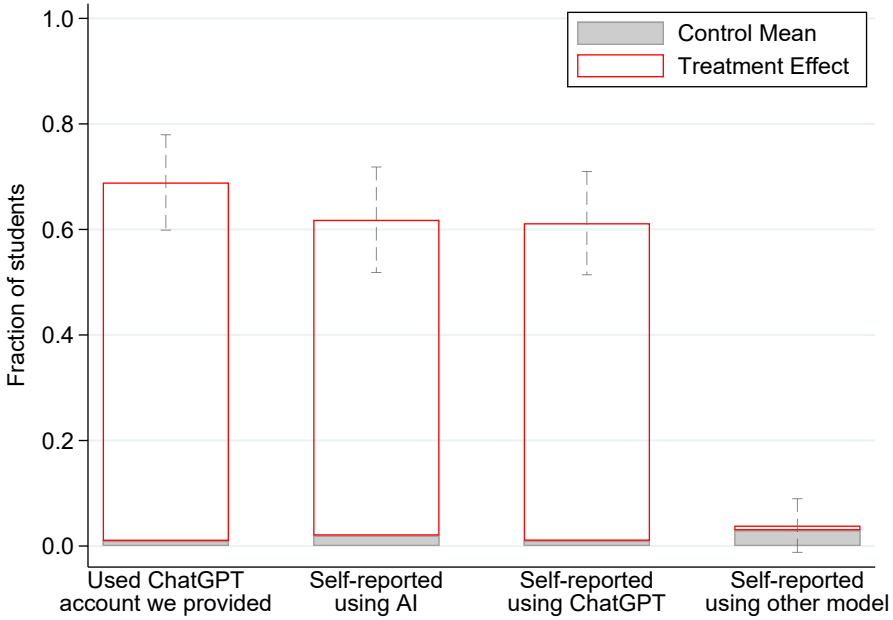
Figures and Tables

Figure 1: Experimental Sessions Timelines



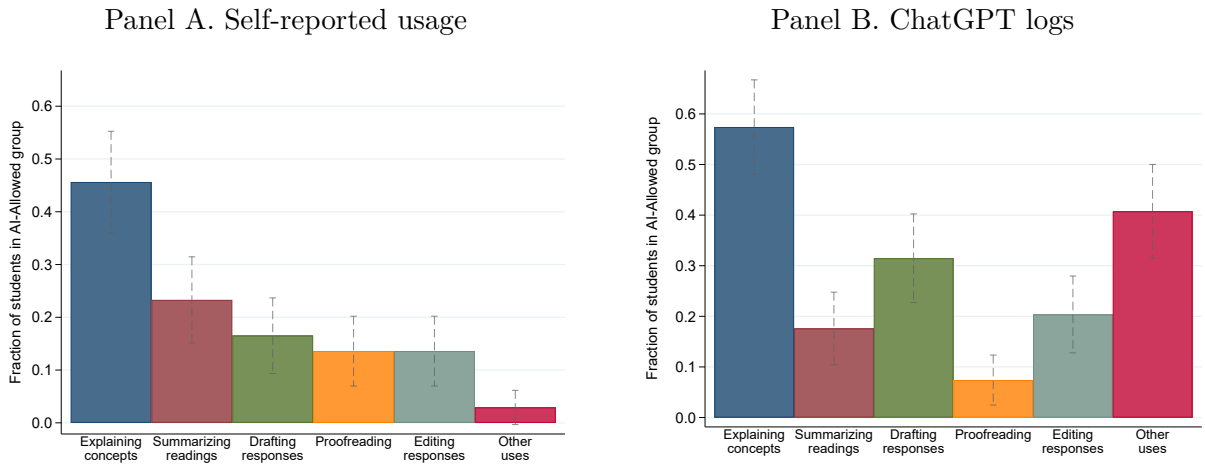
Notes: Panel A shows the structure of Session One, during which participants were randomly assigned to AI-allowed or AI-forbidden conditions during the Learning Phase. Panel B shows Session Two, conducted approximately one week later (mean = 6.9 days), during which all participants completed tasks without AI access. The red dashed box in Panel B indicates that the order of the Endline Test and the Essay was randomized across participants. The Baseline Test and Second Test in Session One each contained 5 multiple-choice questions. The Endline Test in Session Two contained 10 multiple-choice questions.

Figure 2: Impact of AI-Allowance on Generative AI Usage During the Learning Phase



Notes: This figure shows the impact of AI access on generative AI usage during the learning phase in Session One. Gray bars represent control group means, while red outlines show treatment effects. The first measure combines ChatGPT account usage data with direct proctor observations. The remaining three measures are based on self-reported usage collected in the exit survey. Error bars represent 95 percent confidence intervals.

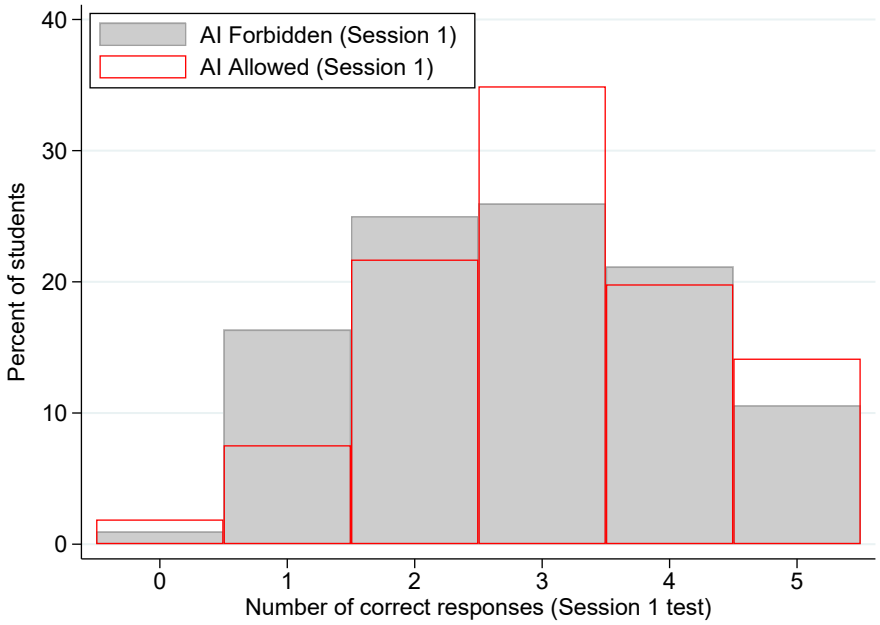
Figure 3: Types of Generative AI Use During the Learning Phase



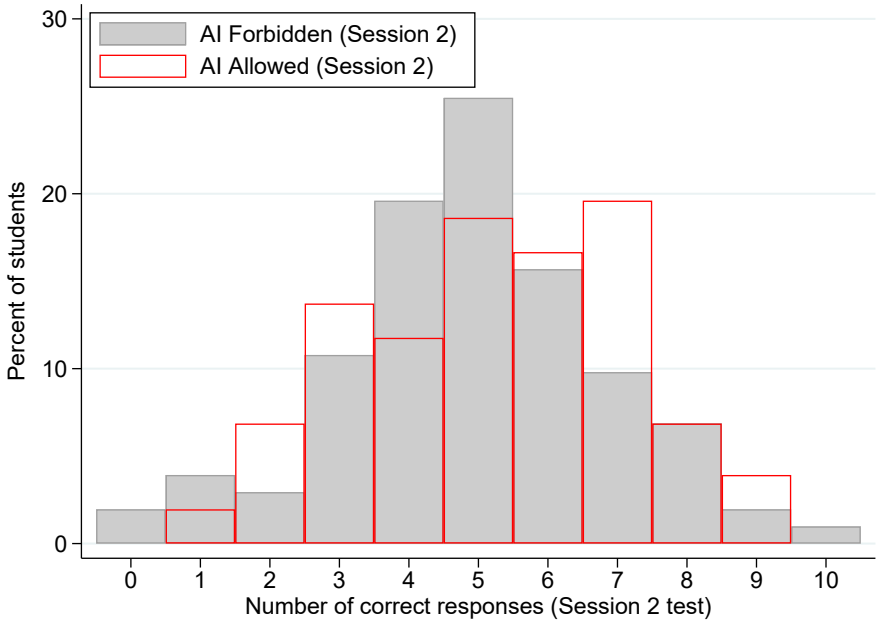
Notes: This figure shows AI usage types among participants in the AI-allowed condition during Session One. Both panels report the fraction of treated students whose AI use included each category; bars are directly comparable across panels. Panel A presents self-reported usage types collected in the post-experiment survey; participants could select multiple categories. Panel B is constructed from the actual ChatGPT conversation logs: we first classify each student prompt into one of six categories (explaining concepts, summarizing readings, drafting responses, proofreading, editing responses, and other uses) using Claude Opus, then collapse to the student level by flagging whether any of a student's prompts fell into each category. Students who were assigned to the AI-allowed condition but did not use the provided ChatGPT account are coded as zero for all categories, so both panels share the same denominator (all treated students).

Figure 4: Distribution of Test Performance by Treatment Group

Panel A. Session One (Knowledge Acquisition)

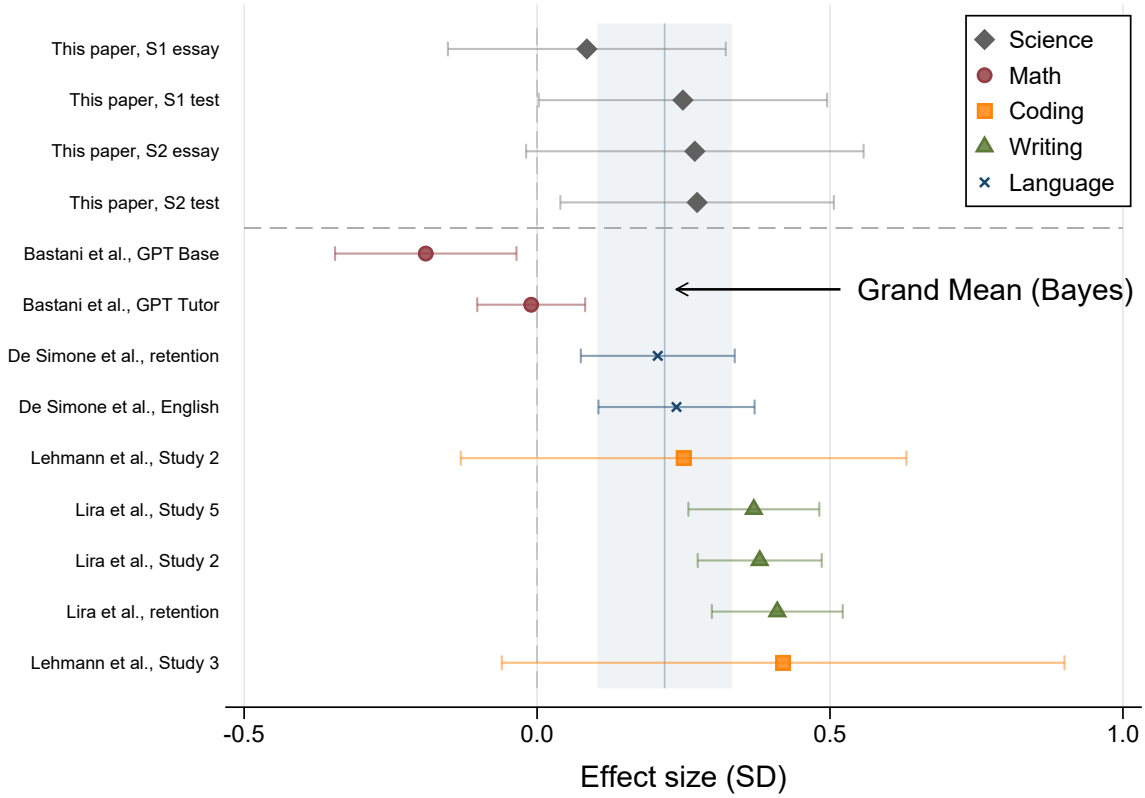


Panel B. Session Two (Knowledge Retention)



Notes: This figure shows the distribution of test scores by treatment assignment. The x-axis represents the percent of correct responses on the knowledge test. The y-axis shows the fraction of students achieving each score. Gray bars represent students in the AI Forbidden group, while red outlined bars represent students in the AI Allowed group. Panel A shows the distribution of scores on the immediate test (5 questions) administered at the end of Session One. Panel B shows the distribution of scores on the retention test (10 questions) administered in Session Two, approximately one week after the learning phase; scores are binned into 20-percentage-point intervals.

Figure 5: Effect Sizes Across AI-and-Learning Experiments

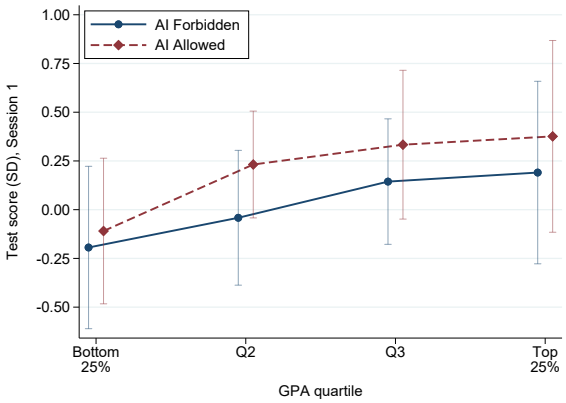


Notes: This figure compares treatment effect sizes (in standard deviations) across experiments that provide students with AI tools during a learning task and measure performance on unassisted assessments. Marker shapes indicate the outcome domain: diamonds for science (this paper), circles for math, squares for coding, triangles for writing, and crosses for language learning. Our estimates include both knowledge test scores and essay quality for Session One (immediate) and Session Two (retention), shown above the dashed separator line. Literature estimates are from Bastani et al. (2025) (math, high school students in Turkey), De Simone et al. (2025) (English language, primary school students in Nigeria), Lehmann et al. (2025) (Python coding, university students in the Netherlands), and Lira et al. (2025) (professional writing, online participants). Horizontal lines show 95% confidence intervals. The shaded band and vertical line show the random-effects grand mean and its 95% confidence interval, estimated via the DerSimonian and Laird (1986) method. A dashed vertical line marks zero.

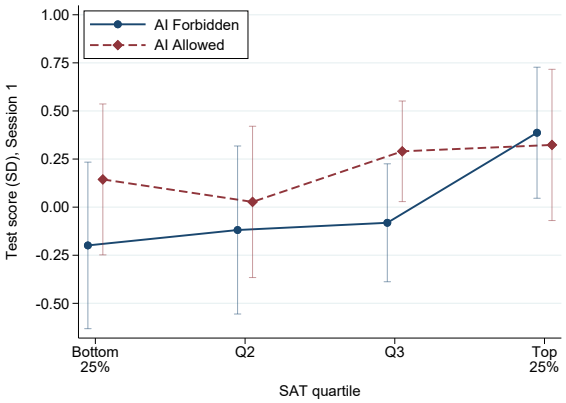
Figure 6: Treatment Effects on Test Scores by GPA and SAT Quartile

Session One

Panel A. By GPA quartile

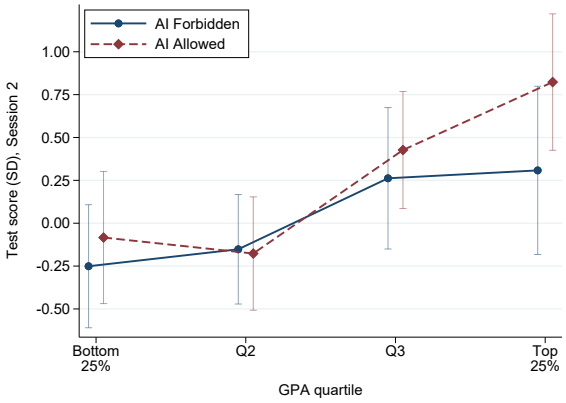


Panel B. By SAT quartile

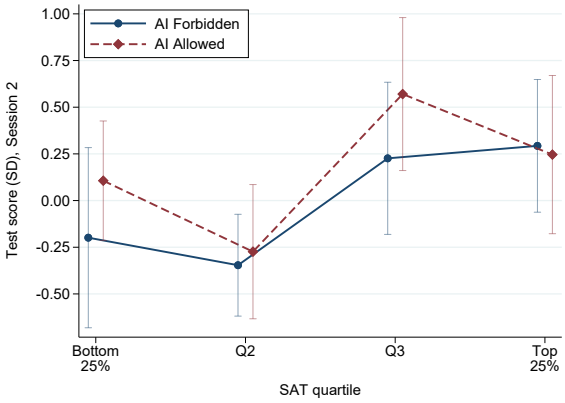


Session Two

Panel C. By GPA quartile

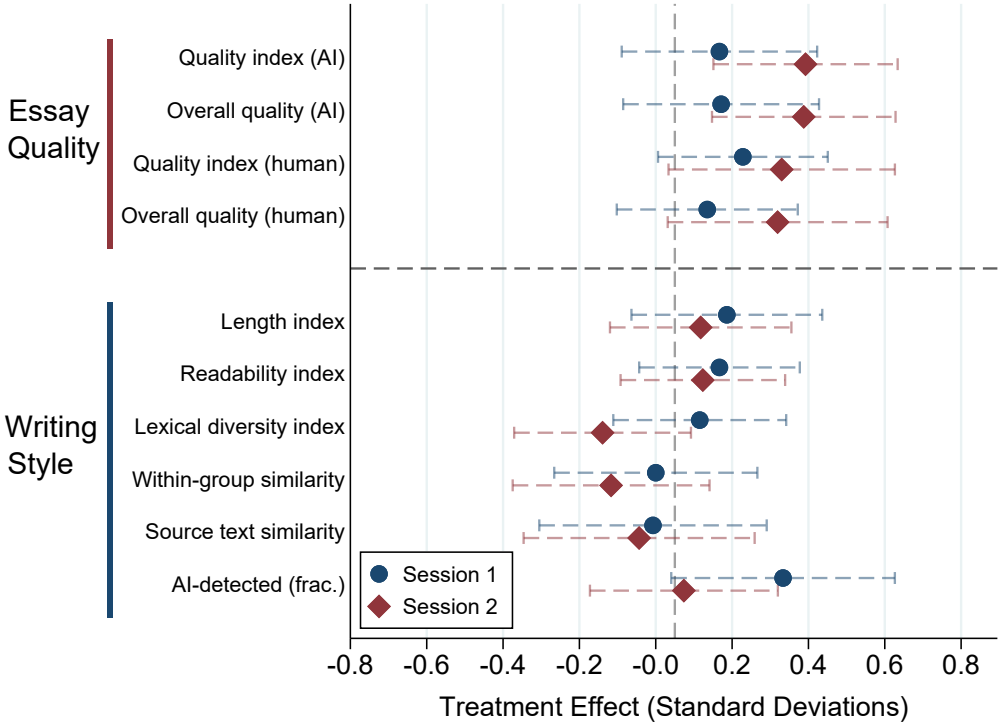


Panel D. By SAT quartile



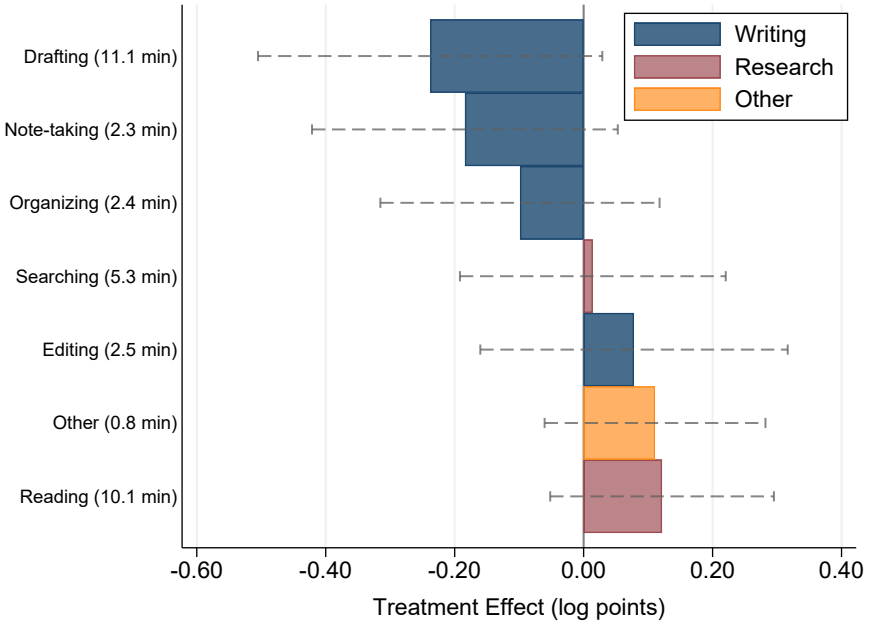
Notes: This figure shows mean test scores by treatment status across quartiles of GPA (Panels A and C) and SAT scores (Panels B and D). SAT scores are imputed for students who did not report them using predicted values from a regression on observable characteristics. Connected lines show group means; error bars denote 95% confidence intervals. Test scores are standardized to have mean 0 and standard deviation 1 in the control group.

Figure 7: Effects of AI Access on Essay Quality and Characteristics



Notes: This figure presents treatment effects of AI access on essay quality measures (top panel) and writing style indices (bottom panel), measured in standard deviations. The essay quality measures include overall quality (the grader's holistic assessment) and the quality index (averaging all six rubric dimensions), reported separately for human and AI graders. The writing style indices are: the length index (averaging standardized number of tokens, words, and sentences), the readability index (averaging standardized sentence length, syllables per word, Flesch-Kincaid grade level, and Flesch Reading Ease score, with difficulty measures reversed), the lexical diversity index (averaging standardized type-token ratio and hapax proportion), the homogeneity index (averaging standardized within-group and source text cosine similarity), and the AI-detected index (standardized fraction of text classified as AI-generated by Pangram). Circles represent Session One effects (essays written with or without AI access); diamonds represent Session Two effects (essays written one week later without AI access). Horizontal lines represent 95 percent confidence intervals with standard errors clustered at the individual level. Detailed breakdowns by individual dimension are presented in Appendix Figures A4 and A5.

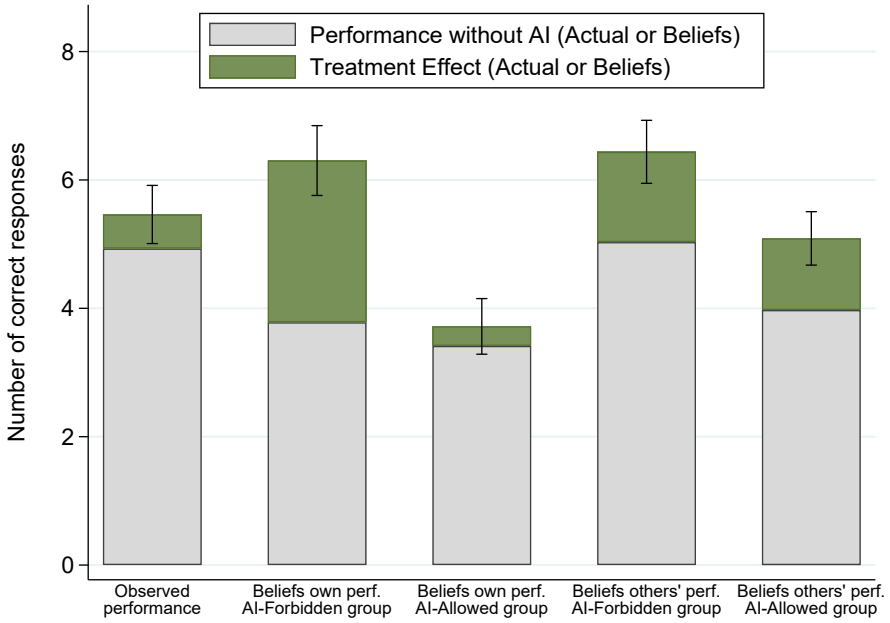
Figure 8: Mechanisms: Time Allocation Across Learning Activities



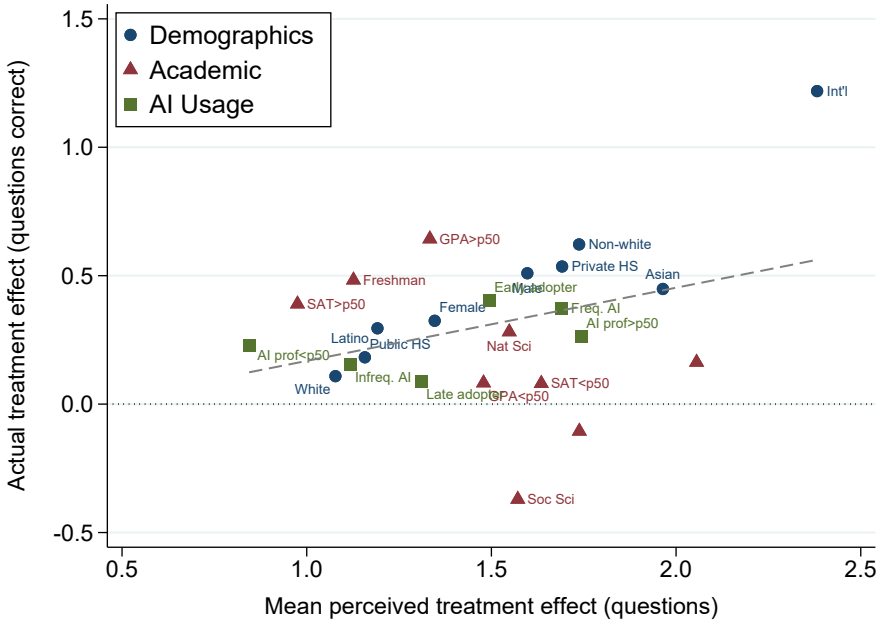
Notes: This figure presents treatment effects of AI access on self-reported time allocation across different learning activities, measured in log minutes. Students reported how many minutes they spent on each activity during the 35-minute learning phase. The survey labels were: "Writing a first draft of your write-up" (Drafting), "Revising and editing your draft" (Editing), "Taking notes on key concepts and facts" (Note-taking), "Planning your writing task structure and arguments" (Organizing), "Finding information about the topic, searching for relevant sources" (Searching), "Reading and comprehending information about the topic" (Reading), and "Casually browsing the web, unrelated to the topic" and "Other activities" combined (Other). Navy bars represent writing activities, red bars represent research activities, and orange bars represent other activities. Bars represent the estimated treatment effect for each activity, with horizontal lines showing 95 percent confidence intervals. Numbers in parentheses indicate the control group mean.

Figure 9: Actual and Perceived Treatment Effects on Test Performance

Panel A. Beliefs vs. Actual Effects



Panel B. Perceived vs. Actual Effects by Subgroup



Notes: Panel A compares the actual treatment effect of AI access on test performance with students' beliefs about this effect. The first bar shows the observed performance: the gray portion represents the control group mean, and the green portion represents the estimated treatment effect. The remaining bars show students' beliefs about performance with and without AI access. For the AI-Forbidden group (bars 2 and 4), beliefs represent predictions about how they would have performed with AI access. For the AI-Allowed group (bars 3 and 5), beliefs represent counterfactual predictions about how they would have performed without AI access. "Own perf." refers to beliefs about one's own performance, while "others' perf." refers to beliefs about other students' performance. Error bars represent 95 percent confidence intervals. The actual treatment effect is estimated using double-lasso with controls, while belief-based treatment effects are computed as the difference between actual and counterfactual beliefs. Panel B plots the mean perceived treatment effect (own performance) against the actual treatment effect for each demographic, academic, and other subgroup with at least 40 observations. Perceived treatment effects are constructed as

Table 1: Summary Statistics

	Completed Intake (1)	Attended Session One (2)	Attended Both Sessions (3)
Panel A. Demographic characteristics			
Age	20.211	20.148	20.132
Male	0.359	0.352	0.353
White	0.496	0.495	0.510
International student	0.270	0.267	0.275
Public high school	0.539	0.538	0.534
Panel B. Academic background			
Took SAT/ACT	0.742	0.762	0.770
SAT score (conditional)	1387.593	1385.566	1386.571
College GPA	3.680	3.684	3.686
Freshman	0.359	0.376	0.382
Already declared major	0.780	0.779	0.777
Natural-science major	0.301	0.310	0.314
Social-science major	0.359	0.352	0.348
Humanities/Arts/Lang. major	0.109	0.105	0.103
Study hours per week	16.929	16.557	16.686
Panel C. Attendance and comprehension			
Attended session 1	0.824	1.000	1.000
Attended session 2	0.797	0.971	1.000
Attended session 2 (cond. S1)	0.971	0.971	1.000
Attended both sessions	0.797	0.971	1.000
Days between sessions	6.951	6.986	6.971
Comprehension score (0-5)	4.768	4.767	4.765
Panel D. Experimental conditions			
Assigned Blockchain topic	0.365	0.367	0.368
Assigned CRISPR topic	0.327	0.329	0.324
Assigned Carbon-capture topic	0.308	0.305	0.309
Baseline self-assessment (0-10)	1.517	1.524	1.520
Fraction correct (baseline)	0.303	0.304	0.310
Shared info between sessions	0.064	0.064	0.064
Looked up info about topic	0.049	0.049	0.049
Studied topic for session 2	0.005	0.005	0.005
Number of students	256	210	204

Notes: Each column reports means for a mutually exclusive group of students: the full intake (column 1), those who attended Session One (column 2), and those who attended both sessions (column 3). Variable definitions appear in Section 4. GPA is on the 0–4 scale; SAT and ACT scores are shown only for students who took the respective test.

Table 2: Balance of Baseline Characteristics by Treatment Assignment

	Treatment group:		Difference:	
	AI Forbidden	AI Allowed	$\hat{\beta}$	(SE)
Panel A. Demographic characteristics				
Age	20.016	20.400	0.387	(0.228)*
Male	0.405	0.315	-0.088	(0.060)
White	0.524	0.469	-0.059	(0.063)
International student	0.286	0.254	-0.030	(0.056)
Public high school	0.540	0.538	-0.004	(0.063)
Panel B. Academic background				
Took SAT/ACT	0.778	0.708	-0.070	(0.054)
SAT score (conditional)	1392.371	1382.554	-8.559	(21.101)
SAT score (incl. predicted)	1391.418	1382.382	-9.391	(18.046)
College GPA	3.694	3.666	-0.025	(0.038)
Freshman	0.357	0.362	0.009	(0.061)
Already declared major	0.817	0.742	-0.076	(0.052)
Natural-science major	0.302	0.300	0.001	(0.058)
Social-science major	0.397	0.323	-0.081	(0.061)
Humanities/Arts/Lang. major	0.111	0.108	-0.000	(0.039)
Study hours per week	16.864	16.992	0.065	(1.189)
Panel C. Attendance and comprehension				
Attended session 1	0.825	0.822	-0.007	(0.048)
Attended session 2	0.810	0.785	-0.022	(0.050)
Attended session 2 (cond. S1)	0.981	0.962	-0.018	(0.022)
Attended both sessions	0.810	0.785	-0.022	(0.050)
Days between sessions	6.975	6.929	-0.030	(0.156)
Comprehension score (0-5)	4.769	4.766	-0.010	(0.065)
Panel D. Experimental conditions				
Assigned Blockchain topic	0.375	0.355	-0.020	(0.066)
Assigned CRISPR topic	0.327	0.327	0.005	(0.065)
Assigned Carbon-capture topic	0.298	0.318	0.015	(0.062)
Baseline self-assessed knowledge (0-10)	1.558	1.477	-0.071	(0.252)
Fraction correct (baseline)	0.306	0.301	-0.006	(0.038)
N (Students)	126	130	256	

Notes: This table shows average baseline characteristics by treatment assignment. Participants in the "AI Forbidden" group were not allowed to use generative AI tools during the learning phase, while participants in the "AI Allowed" group participants were allowed to use generative AI tools. All characteristics are measured prior to treatment assignment. Heteroskedasticity-robust standard errors in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Table 3: The First Stage Impact of Generative AI Access on AI Usage

Outcome	Control mean (1)	Effect ($\hat{\beta}$) (2)
Panel A. Revealed measures		
Used provided ChatGPT account	0.010	0.679***
		(0.046)
Number of prompts	0.000	3.946***
		(0.429)
Number of conversations	0.000	0.751***
		(0.061)
Panel B. Self-reported measures		
Used any AI	0.020	0.599***
		(0.051)
Used ChatGPT	0.010	0.602***
		(0.050)
Used other AI	0.030	0.009
		(0.026)
N	101	204

Notes: Each row reports the effect of being assigned to the *AI-allowed* group on the indicated measure of AI usage. Panel A presents revealed measures: *Used provided ChatGPT account* equals one if at least one prompt was recorded in the ChatGPT account provided to the student or if a proctor directly observed the student using ChatGPT; *Number of prompts* is the total number of prompts sent to the provided ChatGPT account during the session; *Number of conversations* counts the number of distinct ChatGPT conversations (i.e., separate chat threads) initiated during the session. Panel B presents self-reported measures collected in Session Two. All specifications include controls selected through a double-lasso procedure (Belloni et al., 2014). Heteroskedasticity-robust standard errors in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Table 4: The Impact of Generative AI on Learning

	Session One			Session Two		
Outcome	Control mean (1)	ITT (2)	TOT (3)	Control mean (4)	ITT (5)	TOT (6)
Self−assessed	4.788	−0.034	−0.050	3.842	0.109	0.159
		(0.192)	(0.283)		(0.199)	(0.290)
Frac. correct	0.563	0.063**	0.093*	0.493	0.053**	0.078**
		(0.032)	(0.047)		(0.023)	(0.035)
Test score (SD)	0.000	0.249**	0.366*	0.000	0.273**	0.401**
		(0.126)	(0.186)		(0.119)	(0.179)
N	104	208	208	102	204	204

Notes: Each row reports the effect of being assigned to the *AI-allowed* group on the indicated outcome. Self-assessed knowledge is measured on a 0-10 scale where 0 indicates "I know nothing about this topic" and 10 indicates "I am an expert." Fraction correct is measured on a 0-1 scale. Test score (SD) combines performance across all questions and is standardized to have mean 0 and standard deviation 1 in the control group. Columns 1 and 4 report the control group mean. Columns 2 and 5 report intent-to-treat (ITT) estimates. Columns 3 and 6 report treatment-on-the-treated (TOT) estimates from two-stage least squares, instrumenting actual ChatGPT use with random treatment assignment. All specifications include controls selected through a double-lasso procedure (Belloni et al., 2014). Heteroskedasticity-robust standard errors in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Table 5: The Distributional Impact of Generative AI on Learning

Outcome	Session One			Session Two		
	Control mean (1)	ITT (2)	TOT (3)	Control mean (4)	ITT (5)	TOT (6)
Above 20th pctile	0.990	−0.010	−0.014	0.941	0.048*	0.069*
		(0.016)	(0.024)		(0.028)	(0.041)
Above 40th pctile	0.827	0.079	0.116	0.804	−0.001	−0.002
		(0.048)	(0.071)		(0.050)	(0.074)
Above 60th pctile	0.577	0.115*	0.169*	0.353	0.064	0.095
		(0.065)	(0.096)		(0.064)	(0.095)
Above 80th pctile	0.317	0.025	0.038	0.098	0.043	0.062
		(0.061)	(0.093)		(0.041)	(0.061)
Above 100th pctile	0.106	0.037	0.055	0.010	−0.009	−0.013
		(0.044)	(0.064)		(0.009)	(0.013)
N	104	210	210	102	204	204

Notes: Each row reports the effect of being assigned to the *AI-allowed* group on the probability of achieving at least the indicated threshold of correct responses. Columns 1 and 4 report the control group mean. Columns 2 and 5 report intent-to-treat (ITT) estimates. Columns 3 and 6 report treatment-on-the-treated (TOT) estimates from two-stage least squares, instrumenting actual ChatGPT use with random treatment assignment. All specifications include controls selected through a double-lasso procedure (Belloni et al., 2014). Heteroskedasticity-robust standard errors in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Table 6: Effects of Generative AI Access on Essay Quality

Outcome	Session One			Session Two		
	Control mean (1)	ITT (2)	TOT (3)	Control mean (4)	ITT (5)	TOT (6)
Panel A. Overall quality						
Overall quality (human)	6.562	0.162	0.254	5.659	0.579*	0.838*
		(0.230)	(0.353)		(0.316)	(0.474)
Overall quality (AI)	3.892	0.176	0.266	2.714	0.420***	0.617***
		(0.190)	(0.285)		(0.153)	(0.229)
Quality index (human)	6.470	0.297	0.451	5.592	0.570*	0.815*
		(0.189)	(0.278)		(0.308)	(0.460)
Quality index (AI)	4.665	0.161	0.244	3.373	0.385***	0.564***
		(0.180)	(0.272)		(0.138)	(0.208)
Panel B. Individual dimensions (human and AI averaged)						
Accuracy	5.829	0.269**	0.403**	4.626	0.253*	0.369*
		(0.133)	(0.197)		(0.145)	(0.214)
Evidence	5.226	0.119	0.179	3.626	0.168	0.258
		(0.170)	(0.255)		(0.155)	(0.237)
Relevance	5.547	0.319*	0.480*	4.715	0.251	0.377
		(0.167)	(0.251)		(0.158)	(0.240)
Organization	5.217	0.252*	0.385*	4.617	0.270	0.410
		(0.152)	(0.230)		(0.170)	(0.264)
Writing style	5.717	0.416***	0.666***	5.095	0.399**	0.619**
		(0.157)	(0.244)		(0.170)	(0.274)
N	80	164	164	70	150	150

Notes: Each row reports the effect of being assigned to the *AI-allowed* group on the indicated outcome. Each dimension is scored on a 0–10 scale. Panel A reports overall quality and the average of the five sub-component scores (accuracy, evidence, relevance, organization, and writing style), separately for human graders (Prolific workers, 3–4 per essay) and AI grading (Claude Opus, one score per essay). Human grading regressions are at the grader-essay level and include grader and batch fixed effects; AI grading regressions are at the student level. Panel B reports individual dimension scores averaged across human and AI graders. Columns 1–3 report Session One results (when AI-allowed students had access to ChatGPT); columns 4–6 report Session Two results (when all students wrote without AI access). ITT = intent-to-treat; TOT = treatment-on-the-treated (instrumenting ChatGPT use with random assignment). All specifications include controls selected through a double-lasso procedure (Belloni et al., 2014). Standard errors clustered at the student level in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Table 7: Effects of Generative AI Access on Essay Characteristics

Outcome	Session One (with AI)			Session Two (No AI)		
	Control mean (1)	ITT (2)	TOT (3)	Control mean (4)	ITT (5)	TOT (6)
Panel A. Writing style						
Length index	0.000	0.132	0.198	0.054	0.063	0.093
		(0.123)	(0.186)		(0.112)	(0.166)
Readability index	0.000	0.092	0.136	0.000	0.060	0.088
		(0.084)	(0.124)		(0.089)	(0.132)
Lex. diversity index	0.000	0.065	0.100	0.000	−0.188	−0.281
		(0.115)	(0.176)		(0.117)	(0.175)
Panel B. Homogeneity and similarity						
Within−group sim.	0.790	−0.002	−0.004	0.761	−0.018	−0.027
		(0.007)	(0.010)		(0.014)	(0.021)
Source text sim.	0.750	−0.004	−0.006	0.705	−0.009	−0.014
		(0.010)	(0.016)		(0.016)	(0.024)
Panel C. AI detection and plagiarism						
AI−detected (frac.)	0.128	0.094*	0.143*	0.041	0.005	0.007
		(0.050)	(0.075)		(0.025)	(0.038)
Plagiarism (frac.)	0.000	0.025	0.036	0.007	0.034	0.049
		(0.025)	(0.037)		(0.041)	(0.060)
N	79	152	152	79	152	152

Notes: Each row reports the effect of being assigned to the *AI-allowed* group on the indicated outcome. Panel A reports z-scored writing style indices, standardized relative to the control group. Length index averages the z-scores of tokens, words, and sentences. Readability index averages four z-scored measures (sentence length, syllables per word, Flesch-Kincaid grade level, and Flesch Reading Ease), with signs oriented so that higher values indicate easier readability. Lexical diversity index averages the z-scored Type-Token Ratio and hapax proportion. Panel B reports textual similarity measures: within-group similarity is the average pairwise cosine similarity between a student's essay embedding and all other essays in the same treatment × topic × prompt cell; source text similarity is the cosine similarity between a student's essay embedding and the embedding of the provided reading material. Both are computed using the `all-mpnet-base-v2` sentence-transformer model. Panel C reports AI detection and plagiarism measures: AI-detected (frac.) is the fraction of text classified as AI-generated by the Pangram AI content detector; Plagiarism (frac.) is the fraction of text flagged as plagiarized by Pangram's plagiarism checker. Columns 1 and 4 report the control group mean. Columns 2 and 5 report intent-to-treat (ITT) estimates. Columns 3 and 6 report treatment-on-the-treated (TOT) estimates from two-stage least squares, instrumenting actual ChatGPT use with random treatment assignment. All specifications include controls selected through a double-lasso procedure (Belloni et al., 2014). Heteroskedasticity-robust standard errors in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$. Appendix Table A2 reports results for individual writing characteristics.

Table 8: Heterogeneity by Type of AI Interaction

	Session One			Session Two		
	Test score (1)	Overall quality (2)	Quality index (3)	Test score (4)	Overall quality (5)	Quality index (6)
Overall (TOT)	0.366*	0.362	0.530***	0.401**	0.387*	0.360*
	(0.186)	(0.225)	(0.197)	(0.179)	(0.228)	(0.210)
Automation users	0.251	0.448*	0.544**	0.193	−0.090	−0.037
	(0.239)	(0.233)	(0.232)	(0.190)	(0.242)	(0.235)
Augmentation users	0.428**	0.210	0.348**	0.303*	0.436**	0.351**
	(0.173)	(0.161)	(0.135)	(0.154)	(0.196)	(0.167)
N	147	115	115	143	103	103

Notes: The first row reports the TOT estimate from the full sample, obtained by instrumenting actual ChatGPT use with random assignment to the AI-Allowed group. The remaining rows compare each user type to the full control group, where user types are classified from ChatGPT conversation logs (see Appendix B.6). By construction, all treated students in these subgroups are compliers. Because students with mixed conversations (containing both automation and augmentation elements) are included in both subgroups, the two rows are not mutually exclusive and their coefficients need not average to the overall effect. Students who did not use ChatGPT or whose conversations were classified as "Other" appear only in the overall row. All specifications use controls selected by double-lasso on the full sample (Belloni et al., 2014), with strata fixed effects and standard errors clustered at the student level. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Table 9: Mechanisms: Engagement, Experience, and Academic Integrity

Outcome	Control mean (1)	ITT (2)	TOT (3)
Panel A. Engagement			
Opened PDF	0.712	0.065	0.095
		(0.060)	(0.089)
Time on task (Qualtrics, min.)	32.611	−0.199	−0.293
		(0.673)	(0.992)
Time on task (self−reported, min.)	34.312	−0.551	−0.822
		(0.671)	(1.001)
Remembers topic	0.960	−0.018	−0.027
		(0.030)	(0.045)
Panel B. Learning experience			
Enjoyed learning (continuous)	5.221	0.676**	1.016**
		(0.307)	(0.471)
Enjoyed learning (above median)	0.337	0.152**	0.225**
		(0.064)	(0.096)
Found effective (continuous)	5.000	0.099	0.147
		(0.293)	(0.434)
Found effective (above median)	0.423	0.052	0.078
		(0.067)	(0.100)
Panel C. Academic integrity			
Caught by proctor	0.029	0.057*	0.084*
		(0.032)	(0.047)
Self−reported AI use	0.020	0.086**	0.126**
		(0.034)	(0.050)
Any integrity violation	0.050	0.126***	0.185***
		(0.044)	(0.064)
N	101	204	204

Notes: Each row reports the effect of being assigned to the *AI-allowed* group on the indicated outcome. Panel A examines engagement: Opened PDF is a binary indicator for clicking the provided reading materials; time measures are in minutes. Panel B examines the learning experience: task enjoyment and perceived effectiveness are measured on 0-10 scales, with binary indicators for above-median values. Panel C examines academic integrity during knowledge tests where all participants were prohibited from using external resources. All specifications include controls selected through a double-lasso procedure (Belloni et al., 2014). Heteroskedasticity-robust standard errors in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

References

Acemoglu, D. and Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2):3–30.

Ammari, T., Chen, M., Zaman, S. M. M., and Garimella, K. (2025). How students (really) use ChatGPT: Uncovering experiences among undergraduate students.

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.

Becker, J., Rush, N., Barnes, E., and Rein, D. (2025). Measuring the impact of early-2025 AI on experienced open-source developer productivity. Technical report, METR.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, 28(2):29–50.

Bettinger, E. P., Fox, L., Loeb, S., and Taylor, E. S. (2017). Virtual classrooms: How online college courses affect student success. *American Economic Review*, 107(9):2855–2875.

Bick, A., Blandin, A., and Deming, D. J. (2026). The Rapid Adoption of Generative AI. *Management Science*.

Brynjolfsson, E., Li, D., and Raymond, L. (2025). Generative AI at Work. *The Quarterly Journal of Economics*, 140(2):889–942.

Brynjolfsson, E. and Mitchell, T. (2017). What can machine learning do? workforce implications. *Science*, 358(6370):1530–1534.

Brysbaert, M. (2019). How many words do we read per minute? a review and meta-analysis of reading rate. *Journal of Memory and Language*, 109:104047.

Budzyń, K., Romańczyk, M., Kitala, D., Kołodziej, P., Bugajski, M., Adami, H. O., Blom, J., Buszkiewicz, M., Halvorsen, N., Hassan, C., Romańczyk, T., Holme, Ø., Jarus, K., Fielding, S., Kunar, M., Pellise, M., Pilonis, N., Kamiński, M. F., Kalager, M., Bretthauer, M., and Mori, Y. (2025). Endoscopist Deskilling Risk After Exposure to Artificial Intelligence in Colonoscopy: A Multicentre, Observational Study. *The Lancet Gastroenterology & Hepatology*, 10(10):896–903.

Bulman, G. and Fairlie, R. W. (2016). Technology and education: Computers, software, and the internet. In Hanushek, E. A., Machin, S. J., and Woessmann, L., editors, *Handbook of the Economics of Education*, volume 5, pages 239–280. Elsevier.

Caldwell, J. E. (2007). Clickers in the large classroom: Current research and best-practice tips. *CBE—Life Sciences Education*, 6(1):9–20.

Carter, S. P., Greenberg, K., and Walker, M. S. (2017). The impact of computer usage on academic performance: Evidence from a randomized trial at the United States Military Academy. *Economics of Education Review*, 56:118–132.

Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading*, 36(2):84–95.

Chiang, C.-H. and Lee, H.-y. (2023). Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 15607–15631. Association for Computational Linguistics.

College Board (2025). SAT nationally representative and user percentiles. https://research.collegeboard.org/reports/sat-suite/understanding-scores/sat. Accessed: 2025-10-31.

Contractor, Z. and Reyes, G. (2025). Generative AI in higher education: Evidence from an elite college. IZA Discussion Paper Nr. 18055.

Cristia, J., Ibarrarán, P., Cueto, S., Santiago, A., and Severín, E. (2017). Technology and child development: Evidence from the One Laptop per Child program. *American Economic Journal: Applied Economics*, 9(3):295–320.

Cui, Z. K., Demirer, M., Jaffe, S., Musolff, L., Peng, S., and Salz, T. (2026). The Effects of Generative AI on High Skilled Work: Evidence from Three Field Experiments with Software Developers. *Management Science*.

De Simone, G., Ferroni, M., Ferroni, S., and Ferroni, N. (2025). From chalkboards to chatbots: The AI tutoring revolution in Italian secondary schools. Technical report, IZA Discussion Paper.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188.

Dettling, L. J., Goodman, S., and Smith, J. (2018). Every little bit counts: The impact of high-speed internet on the transition to college. *The Review of Economics and Statistics*, 100(2):260–273.

Emi, B. (2024). Technical report on the Pangram AI-generated text classifier. Technical report.

Escueta, M., Nickow, A. J., Oreopoulos, P., and Quan, V. (2020). Upgrading education with technology: Insights from experimental research. *Journal of Economic Literature*, 58(4):897–996.

Figlio, D. N., Rush, M., and Yin, L. (2013). Is it live or is it internet? Experimental estimates of the effects of online instruction on student learning. *Journal of Labor Economics*, 31(4):763–784.

Handa, K., Bent, D., Tamkin, A., McCain, M., Durmus, E., Stern, M., Schiraldi, M., Huang, S., Ritchie, S., Syverud, S., Jagadish, K., Vo, M., Bell, M., and Ganguli, D. (2025). Anthropic education report: How university students use Claude.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334.

Jabarian, B. and Imas, A. (2025). Artificial writing and automated detection. Technical report, Becker Friedman Institute, University of Chicago.

Jensen, R. (2010). The (perceived) returns to education and the demand for schooling. *The Quarterly Journal of Economics*, 125(2):515–548.

Kestin, G., Miller, K., Klales, A., Milbourne, T., and Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15:17458.

Kim, J., Lee, H., Park, J., and Koedinger, K. R. (2025). Generative AI can improve performance and engagement without harming learning. Technical report, KAIST.

Kreijkes, P., van der Spoel, I., and van der Schaaf, M. (2025). The effects of using ChatGPT as a learning tool on learning outcomes. *Computers and Education: Artificial Intelligence*, 8:100340.

Kumar, S., Mannekote, A., Tong, M., Hegde, S., Rosenzweig, E. Q., and Fox, J. M. (2023). Math education with large language models: Peril or promise? Technical report.

LearnLM Team (2025). LearnLM: Improving Gemini for Learning. Google DeepMind Technical Report.

Lee, H.-P. H., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., and Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM.

Lehmann, M., Cornelius, P. B., and Sting, F. J. (2025). AI meets the classroom: When do large language models harm learning?

Lewin, C., Somekh, B., and Steadman, S. (2008). Embedding interactive whiteboards in teaching and learning: The process of change in pedagogic practice. *Education and Information Technologies*, 13(4):291–303.

Lira, B., Rogers, T., Goldstein, D. G., Ungar, L., and Duckworth, A. L. (2025). Coach not crutch: Evidence that AI can improve writing skill despite reducing effort. Technical report, University of Pennsylvania. arXiv:2502.02880.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023). G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. Association for Computational Linguistics.

Malamud, O. and Pop-Eleches, C. (2011). Home computer use and the development of human capital. *The Quarterly Journal of Economics*, 126(2):987–1027.

Masrour, E., Emi, B. N., and Spero, M. (2025). DAMAGE: Detecting adversarially modified AI generated text. In *Workshop on Detecting AI Generated Content (GenAIDetect), COLING*.

Mozannar, H., Bansal, G., Fourney, A., and Horvitz, E. (2024). Reading between the lines: Modeling user behavior and costs in AI-assisted programming. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM.

Nam, J. (2023). 56% of college students have used AI on assignments or exams. BestColleges. Edited by Lyss Welding. Accessed June 11, 2025.

Noy, S. and Zhang, W. (2023). Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. *Science*, 381(6654):187–192.

OpenAI (2025). Building an AI-ready workforce: A look at college student ChatGPT adoption in the US. Technical report, OpenAI.

Peng, S., Kalliamvakou, E., Cihon, P., and Demirer, M. (2023). The Impact of AI on Developer Productivity: Evidence from GitHub Copilot.

Ravšelj, D., Keržič, D., Tomaževič, N., Umek, L., Brezovar, N., et al. (2025). Higher Education Students' Perceptions of ChatGPT: A Global Study of Early Reactions. *PLOS ONE*, 20(2):e0315011.

Stöhr, C., Ou, A. W., and Malmström, H. (2024). Perceptions and Usage of AI Chatbots Among Students in Higher Education Across Genders, Academic Levels and Fields of Study. *Computers and Education: Artificial Intelligence*, 7:100259.

Thai, K., Emi, B., Masrour, E., and Iyyer, M. (2026). EditLens: Quantifying the extent of AI editing in text. In *International Conference on Learning Representations (ICLR)*.

Vigdor, J. L., Ladd, H. F., and Martinez, E. (2014). Scaling the digital divide: Home computer technology and student achievement. *Economic Inquiry*, 52(3):1103–1119.

Wiswall, M. and Zafar, B. (2015). Determinants of college major choice: Identification using an information experiment. *The Review of Economic Studies*, 82(2):791–824.

Xu, Y.-F., Lew, M. D. N., Yeo, M., and Chan, K. K. H. (2025). Metacognitive support in generative AI environments: Effects on learning performance and prompt quality. *Computers and Education: Artificial Intelligence*, 8:100351.

Yakura, H., Lopez-Lopez, E., Brinkmann, L., Serna, I., Gupta, P., Soraperra, I., and Rahwan, I. (2025). Empirical evidence of large language model's influence on human spoken communication.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36.

Appendix

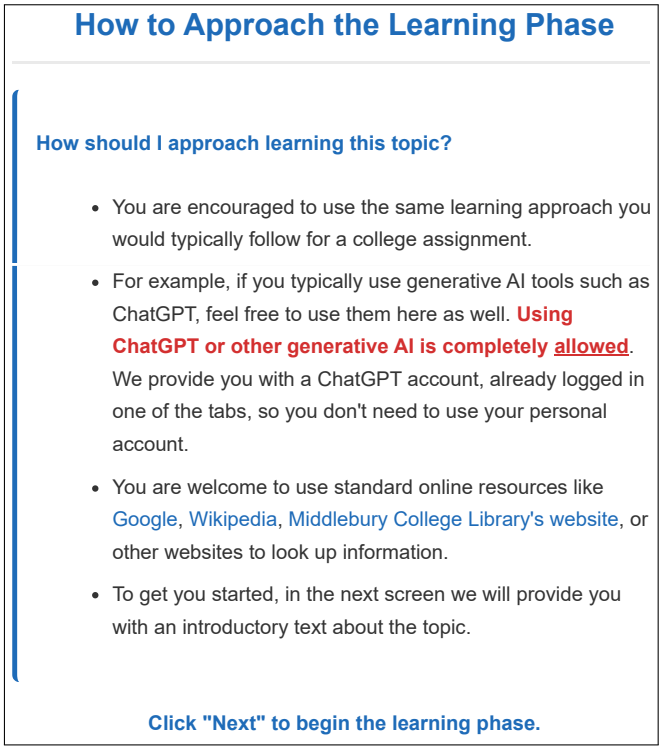
A Appendix Figures and Tables

Figure A1: Computer Lab Setup with Privacy Dividers

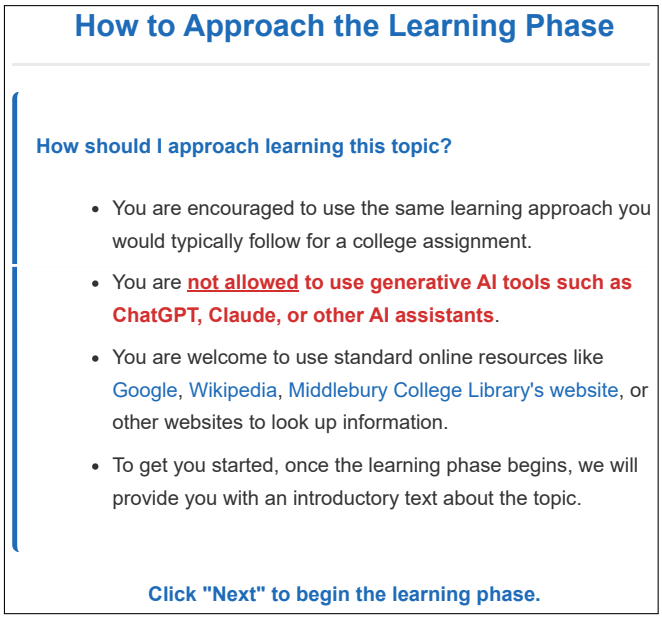
Note: This figure shows the computer lab setup used during experimental sessions. Each workstation was equipped with privacy dividers to minimize distractions and prevent participants from viewing other screens.

Figure A2: Treatment Instructions Displayed to Participants

Panel A. AI-Allowed Condition



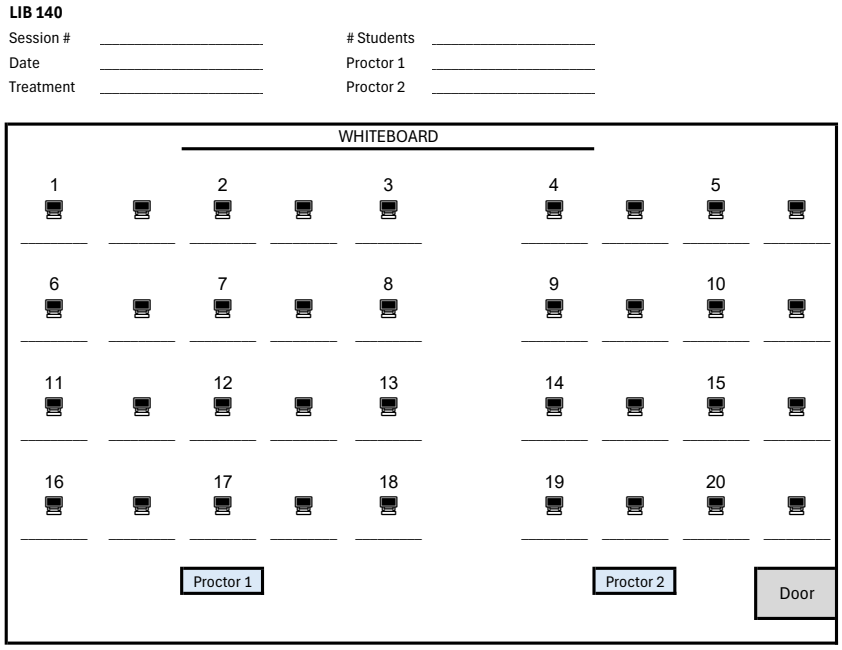
Panel B. AI-Forbidden Condition



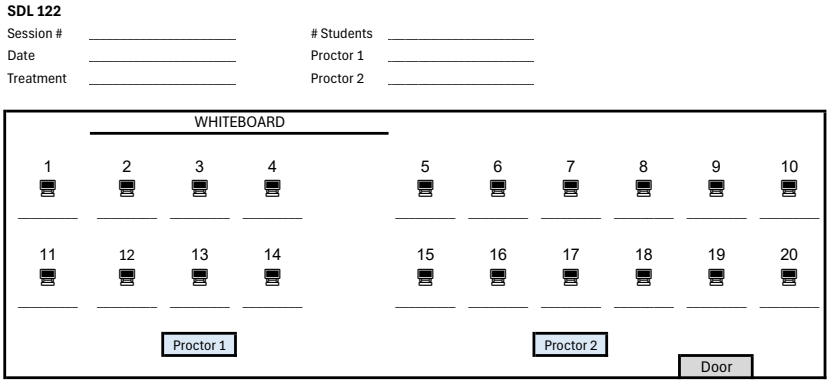
Note: This figure shows the treatment-specific instructions displayed to participants in the survey interface. Panel A shows the message for participants randomly assigned to the AI-allowed condition, who could use generative AI tools during the 35-minute learning phase. Panel B shows the message for participants in the AI-forbidden condition, who were prohibited from using AI tools.

Figure A3: Laboratory Seating Charts

Panel A. Library Computer Lab (LIB 140)

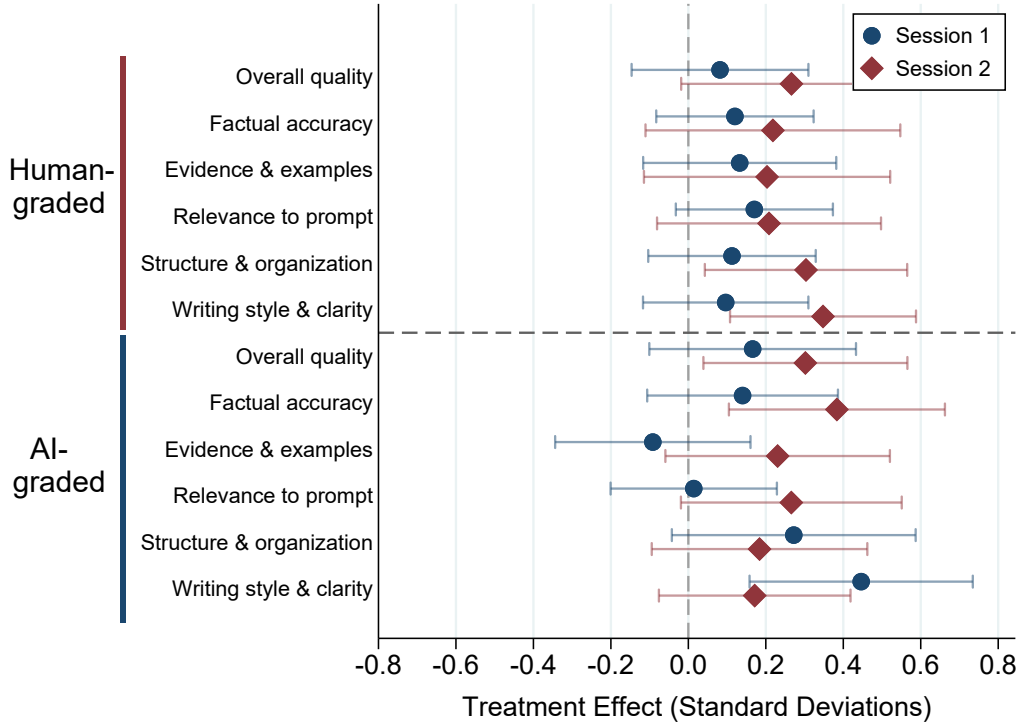


Panel B. Language Building Computer Lab (SDL 122)



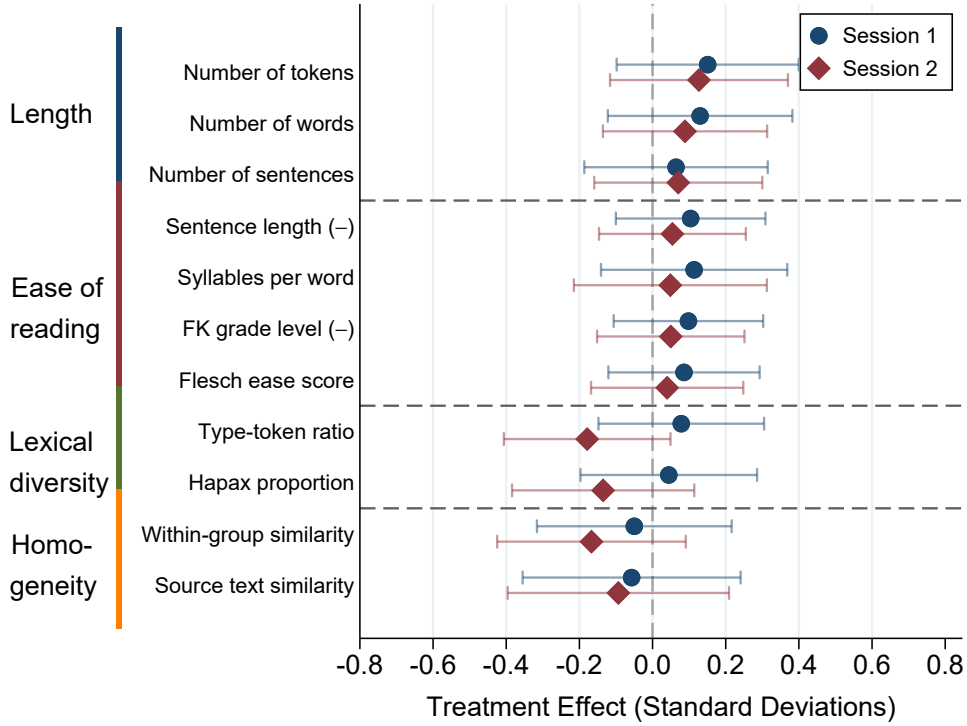
Note: This figure shows the seating charts used by laboratory staff to monitor compliance during experimental sessions. Each numbered square represents a computer workstation. For each session time slot, one lab was randomly assigned to the AI-allowed condition and the other to the AI-forbidden condition. Proctors used these charts to record any instances of unauthorized resource use, with two proctors assigned to each laboratory. The physical separation of treatment conditions across different buildings prevented cross-contamination between experimental groups.

Figure A4: Effects of AI Access on Essay Quality, by Dimension



Notes: This figure presents treatment effects of AI access on individual essay quality dimensions, measured in standard deviations. The top panel shows human-graded dimensions: essays were evaluated by independent graders on six dimensions using a standardized rubric (overall quality, factual accuracy, use of evidence and examples, relevance to the prompt, structure and organization, and writing style and clarity). Human-graded regressions include grader and batch fixed effects. The bottom panel shows the same six dimensions as scored by an AI grader (Claude). Each dimension is scored on a scale from 1 to 10 and standardized to have mean 0 and standard deviation 1 in the control group. Circles represent Session One effects (essays written with or without AI access); diamonds represent Session Two effects (essays written one week later without AI access). Horizontal lines represent 95 percent confidence intervals with standard errors clustered at the individual level.

Figure A5: Effects of AI Access on Essay Characteristics, by Dimension

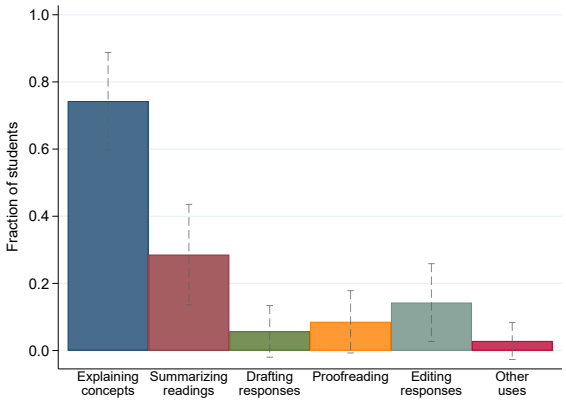


Notes: This figure presents treatment effects of AI access on individual essay characteristics, measured in standard deviations. Each variable is standardized to have mean zero and standard deviation one in the control group. Variables are grouped into four categories, indicated by colored vertical bars: length (number of tokens, words, and sentences), ease of readability (sentence length, syllables per word, Flesch-Kincaid grade level, and Flesch Reading Ease score, with difficulty measures reversed so that higher values indicate easier readability), lexical diversity (type-token ratio and hapax proportion), and homogeneity (within-group cosine similarity and source text cosine similarity). Circles represent Session One effects; diamonds represent Session Two effects. Horizontal lines represent 95 percent confidence intervals with standard errors clustered at the individual level.

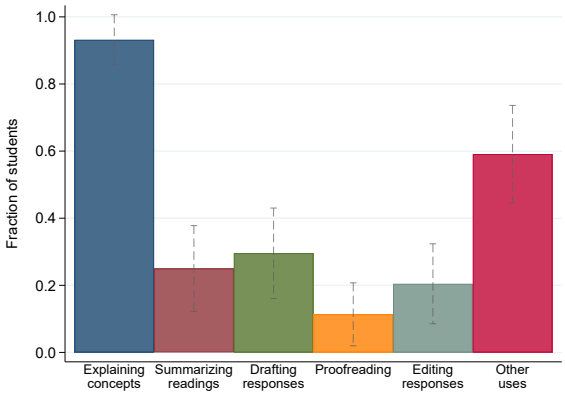
Figure A6: Types of Generative AI Use by Automation and Augmentation Users

Augmentation users

Panel A. Self-reported usage

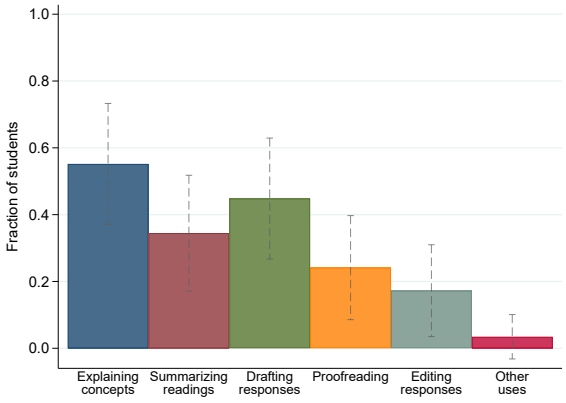


Panel B. Revealed usage (ChatGPT logs)

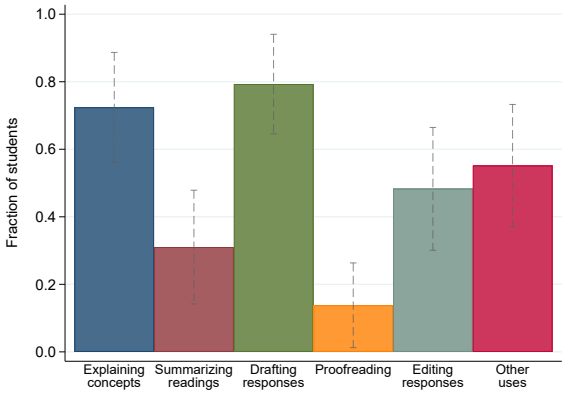


Automation users

Panel C. Self-reported usage



Panel D. Revealed usage (ChatGPT logs)



Notes: This figure replicates the analysis from Figure 3, separately for augmentation and automation users as classified by GPT-4o (see Appendix B.6). Panels A and C present self-reported usage types from the post-experiment survey; participants could select multiple categories. Panels B and D show the fraction of students with at least one prompt classified into each category from the actual ChatGPT conversation logs. Mixed users (with both augmentation and automation conversations) are included in both groups. Error bars denote 95% confidence intervals.

Table A1: Determinants of AI Use Among AI-Allowed Students

	Bivariate regression		Multivariate regression	
	Used account (1)	Self-reported (2)	Used account (3)	Self-reported (4)
Panel A. Demographic characteristics				
Male	0.128	0.113	0.095	0.060
	(0.103)	(0.110)	(0.100)	(0.104)
White	−0.198*	−0.065	−0.179	0.010
	(0.107)	(0.112)	(0.116)	(0.130)
International student	0.154	0.093	0.165	0.044
	(0.115)	(0.127)	(0.150)	(0.162)
Private high school	−0.050	−0.083	−0.176*	−0.142
	(0.096)	(0.106)	(0.106)	(0.122)
Panel B. Academic background				
Natural Sciences major	−0.103	−0.167	−0.131	−0.089
	(0.099)	(0.106)	(0.130)	(0.125)
Social Sciences major	0.125	0.258**	0.077	0.264**
	(0.098)	(0.101)	(0.122)	(0.129)
High GPA (above median)	−0.154*	−0.101	−0.077	−0.052
	(0.089)	(0.100)	(0.105)	(0.123)
Freshman	−0.117	−0.071	−0.234*	−0.014
	(0.098)	(0.109)	(0.136)	(0.141)
Sophomore	0.118	0.119	−0.095	0.043
	(0.101)	(0.125)	(0.138)	(0.180)
Junior	−0.076	−0.016	−0.162	−0.098
	(0.128)	(0.133)	(0.142)	(0.159)
Panel C. AI experience and beliefs				
Frequent AI user (above median)	0.288***	0.251**	0.291***	0.173
	(0.088)	(0.098)	(0.099)	(0.115)
Early AI adopter (pre−Fall 2024)	0.159*	0.161	−0.007	0.021
	(0.094)	(0.103)	(0.112)	(0.131)
High AI proficiency (above median)	0.197**	0.208**	0.020	0.130
	(0.088)	(0.101)	(0.106)	(0.125)
High perceived AI effect (above median)	0.150	0.226**	0.066	0.208*
	(0.090)	(0.100)	(0.096)	(0.106)
N (Students)	106	102	106	102

Notes: Each row reports the association between the indicated characteristic and AI use among AI-allowed students. Columns 1–2 are bivariate; columns 3–4 include all characteristics simultaneously. "Used account" (columns 1, 3) is based on ChatGPT server logs; "Self-reported" (columns 2, 4) is collected in Session Two. All variables are indicators. High GPA is above the sample median. Frequent AI user reports more than occasional AI use during the semester. Early adopter first used AI for academics before Fall 2024. High AI proficiency exceeds the median self-assessed proficiency. High perceived AI effect indicates an above-median perceived treatment effect on test performance (believed score with AI minus believed counterfactual). All specifications control for strata, topic, task version, and lab fixed effects. Standard errors clustered at the student level in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Table A2: Effects of Generative AI Access on Individual Essay Characteristics

	Session One			Session Two		
Outcome	Control mean (1)	ITT (2)	TOT (3)	Control mean (4)	ITT (5)	TOT (6)
Panel A. Length						
Essay is empty	0.000	0.009	0.014	0.039	0.009	0.013
		(0.009)	(0.014)		(0.017)	(0.025)
Tokens	202.692	14.867	22.299	150.724	8.653	12.768
		(12.527)	(18.817)		(8.462)	(12.496)
Words	347.558	20.363	30.409	265.373	11.999	17.302
		(20.159)	(30.182)		(15.494)	(22.367)
Sentences	16.144	0.492	0.738	11.951	0.425	0.644
		(0.981)	(1.469)		(0.708)	(1.077)
Panel B. Readability						
Sentence length	27.091	−4.596	−6.793	25.545	−1.784	−2.630
		(4.604)	(6.817)		(3.365)	(4.964)
Syllables/word	1.776	0.014	0.020	1.714	0.006	0.009
		(0.015)	(0.023)		(0.017)	(0.025)
FK grade level	15.930	−1.682	−2.486	14.598	−0.622	−0.917
		(1.787)	(2.647)		(1.284)	(1.893)
Flesch ease	29.108	3.872	5.724	35.901	1.284	1.893
		(4.750)	(7.036)		(3.410)	(5.028)
Panel C. Lexical diversity						
TTR	0.723	0.007	0.011	0.718	−0.018	−0.026
		(0.010)	(0.016)		(0.011)	(0.017)
Hapax prop.	0.579	0.005	0.007	0.566	−0.018	−0.026
		(0.014)	(0.021)		(0.017)	(0.025)
Panel D. Textual similarity						
Cosine sim.	0.790	−0.002	−0.004	0.761	−0.018	−0.027
		(0.007)	(0.010)		(0.014)	(0.021)
Source text sim.	0.750	−0.004	−0.006	0.705	−0.009	−0.014
		(0.010)	(0.016)		(0.016)	(0.024)
N	79	152	152	79	152	152

Notes: Each row reports the effect of being assigned to the *AI-allowed* group on the indicated outcome. Tokens is the total number of tokens in the essay. Words is the total number of words. Sentences is the total number of sentences. Sentence length is the mean number of words per sentence (sign reversed so higher values indicate shorter sentences). Syllables/word is the mean number of syllables per word. FK grade level is the Flesch-Kincaid Grade Level (sign reversed so higher values indicate simpler text). Flesch ease is the Flesch Reading Ease Score (0-100, higher indicates easier readability). TTR is the Type-Token Ratio. Hapax prop. is the share of words appearing only once. Cosine sim. is the average pairwise cosine similarity within treatment × topic × prompt cells. All specifications include controls selected through a double-lasso procedure (Belloni et al., 2014). Heteroskedasticity-robust standard errors in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Table A3: The Distributional Impact of Generative AI on Essay Quality

Outcome	Session One			Session Two		
	Control mean (1)	ITT (2)	TOT (3)	Control mean (4)	ITT (5)	TOT (6)
Panel A. Human-graded						
Score ≥ 1	0.993	−0.005	−0.007	0.986	0.019**	0.028**
		(0.004)	(0.006)		(0.009)	(0.014)
Score ≥ 3	0.967	0.013	0.019	0.903	0.056	0.079
		(0.013)	(0.019)		(0.045)	(0.064)
Score ≥ 5	0.839	0.024	0.035	0.753	0.053	0.077
		(0.044)	(0.065)		(0.057)	(0.085)
Score ≥ 7	0.553	0.077	0.113	0.348	0.127**	0.184**
		(0.066)	(0.095)		(0.057)	(0.086)
Score ≥ 9	0.174	0.016	0.025	0.086	0.045*	0.064*
		(0.031)	(0.047)		(0.025)	(0.037)
N	304	577	577	279	513	513
Panel B. AI-graded						
Score ≥ 1	0.990	0.000	0.000	0.990	−0.011	−0.016
		(0.014)	(0.020)		(0.018)	(0.026)
Score ≥ 3	0.814	0.047	0.070	0.510	0.075	0.114
		(0.046)	(0.068)		(0.059)	(0.091)
Score ≥ 5	0.343	0.025	0.037	0.102	0.016	0.024
		(0.060)	(0.089)		(0.042)	(0.063)
Score ≥ 7	0.049	0.035	0.051	0.000	0.011	0.016
		(0.037)	(0.054)		(0.011)	(0.016)
N	102	203	203	98	198	198

Notes: Each row reports the effect of being assigned to the *AI-allowed* group on the probability of achieving at least the indicated essay quality threshold. Quality scores range from 0 (lowest) to 10 (highest). Panel A uses human grades from Prolific raters, with the unit of observation at the grader-essay level; specifications include grader and batch fixed effects. Panel B uses AI grades from Claude Opus, with the unit of observation at the student level. Columns 1 and 4 report the control group mean. Columns 2 and 5 report intent-to-treat (ITT) estimates. Columns 3 and 6 report treatment-on-the-treated (TOT) estimates from two-stage least squares, instrumenting actual ChatGPT use with random treatment assignment. All specifications include strata fixed effects and controls selected through a double-lasso procedure (Belloni et al., 2014). Standard errors clustered at the student level in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Table A4: The Impact of Generative AI on Learning, by Topic

Outcome	Session One			Session Two		
	Control mean (1)	ITT (2)	TOT (3)	Control mean (4)	ITT (5)	TOT (6)
Panel A. Blockchain						
Self−assessed	4.795	0.548*	0.843*	3.895	0.211	0.317
		(0.299)	(0.457)		(0.362)	(0.533)
Frac. correct	0.528	0.054	0.084	0.584	0.026	0.041
		(0.053)	(0.083)		(0.037)	(0.058)
Test score (SD)	−0.139	0.211	0.332	0.469	0.136	0.212
		(0.211)	(0.328)		(0.193)	(0.297)
N	39	77	77	38	75	75
Panel B. CRISPR						
Self−assessed	4.735	−0.082	−0.135	3.606	0.174	0.363
		(0.307)	(0.505)		(0.509)	(1.045)
Frac. correct	0.647	0.087	0.124	0.470	0.062	0.126
		(0.062)	(0.089)		(0.068)	(0.150)
Test score (SD)	0.330	0.344	0.488	−0.121	0.319	0.650
		(0.245)	(0.352)		(0.348)	(0.775)
N	34	69	69	33	66	66
Panel C. Carbon capture						
Self−assessed	4.839	0.087	0.130	4.033	−0.030	−0.047
		(0.405)	(0.603)		(0.445)	(0.705)
Frac. correct	0.516	0.098	0.147	0.406	0.024	0.034
		(0.062)	(0.095)		(0.040)	(0.056)
Test score (SD)	−0.187	0.386	0.579	−0.447	0.121	0.176
		(0.246)	(0.374)		(0.207)	(0.291)
N	31	64	64	31	62	62

Notes: This table replicates the main test score results (Table 4) separately by assigned topic. Each panel restricts the sample to students assigned the indicated topic. All specifications mirror those in the main table: ITT estimates from OLS with double-lasso-selected controls (Belloni et al., 2014) and strata fixed effects; TOT estimates instrument ChatGPT use with random assignment. Standard errors clustered at the student level in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Table A5: Robustness of Main Results to Sample Restrictions

		Excluding those who...			
	Baseline	failed attention	failed compreh.	studied between	looked up topic
	(1)	(2)	(3)	(4)	(5)
Panel A. Session One					
Test score (SD)	0.249**	0.391**	0.235	0.245*	0.200
	(0.126)	(0.158)	(0.143)	(0.126)	(0.131)
Overall quality	0.242	0.102	0.121	0.233	0.244
	(0.153)	(0.205)	(0.166)	(0.155)	(0.158)
Quality index	0.349***	0.229	0.186	0.337**	0.355***
	(0.130)	(0.173)	(0.146)	(0.130)	(0.134)
N	208	129	165	207	198
Panel B. Session Two					
Test score (SD)	0.273**	0.305*	0.362***	0.255**	0.237*
	(0.119)	(0.166)	(0.137)	(0.118)	(0.120)
Overall quality	0.253*	0.317*	0.282*	0.274*	0.295*
	(0.148)	(0.188)	(0.160)	(0.150)	(0.160)
Quality index	0.231*	0.314*	0.199	0.242*	0.254*
	(0.136)	(0.187)	(0.144)	(0.138)	(0.148)
N	204	127	161	203	194

Notes: Each cell reports the ITT estimate from regressing the row outcome on a treatment indicator. Column 1 is the baseline specification with controls selected by double-lasso (Belloni et al., 2014). Columns 2–5 use the baseline specification but restrict the sample: column 2 excludes participants who failed the attention check; column 3 excludes those who answered fewer than 5 of 5 comprehension questions correctly; column 4 excludes those who reported studying the topic between sessions; column 5 excludes those who reported looking up the topic between sessions. Overall quality is the average of human and AI grader scores on a 0–10 holistic quality dimension. Quality index averages the five non-holistic dimensions (accuracy, evidence, relevance, organization, and writing style), each averaged across human and AI graders. All specifications include strata fixed effects and cluster standard errors at the student level. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Table A6: Heterogeneity of Treatment Effects by Student Characteristics

	Session One			Session Two		
	Test score (1)	Overall quality (2)	Quality index (3)	Test score (4)	Overall quality (5)	Quality index (6)
Overall (TOT)	0.366*	0.362	0.530***	0.401**	0.387*	0.360*
	(0.186)	(0.225)	(0.197)	(0.179)	(0.228)	(0.210)
Treated ×						
Female	0.441*	−0.305	−0.227	−0.088	0.071	−0.135
	(0.259)	(0.327)	(0.296)	(0.257)	(0.346)	(0.323)
International	0.277	0.586*	0.317	0.475*	−0.006	0.031
	(0.296)	(0.334)	(0.311)	(0.285)	(0.336)	(0.325)
Nonwhite	0.351	0.067	0.151	0.199	−0.073	−0.157
	(0.264)	(0.326)	(0.275)	(0.248)	(0.324)	(0.305)
STEM	0.229	−0.421	−0.143	−0.109	−0.031	−0.186
	(0.263)	(0.323)	(0.300)	(0.261)	(0.327)	(0.303)
Frequent AI user	0.585**	0.459	0.406	0.123	−0.099	0.072
	(0.268)	(0.304)	(0.267)	(0.229)	(0.299)	(0.272)
Early AI adopter	0.031	0.201	0.341	0.163	−0.244	−0.112
	(0.266)	(0.333)	(0.272)	(0.235)	(0.307)	(0.285)
High AI proficiency	−0.100	0.436	0.450*	−0.178	0.229	0.234
	(0.259)	(0.302)	(0.263)	(0.236)	(0.322)	(0.287)
N	147	115	115	143	103	103

Notes: The first row ("Overall") shows the main treatment effect on the full sample. The remaining rows report the coefficient on the interaction Treated × Subgroup indicator from a regression that also includes the treatment indicator and the subgroup indicator as main effects. All specifications use controls selected by double-lasso on the full sample (Belloni et al., 2014), with strata fixed effects and standard errors clustered at the student level. Subgroups defined by median splits use the full-sample median. SAT scores are imputed for students who did not report them. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

B Empirical Appendix

B.1 Reading Materials

To facilitate a standardized learning experience across treatment conditions, we developed original instructional texts on each of the three assigned topics: Blockchain, Carbon Capture, and CRISPR. The full text of each document is available in the Supplementary Materials. These materials were intentionally designed to be of comparable complexity and structure, ensuring that all participants face similar cognitive demands regardless of their topic assignment.

Each text follows a parallel structure, beginning with a general overview of the technology, followed by its historical development, technical underpinnings, core applications, and limitations. The materials are information-rich but self-contained, enabling comprehension without prior topic knowledge. Texts range from 1,253 to 1,399 words in length. Based on an average silent reading speed of approximately 250 words per minute for college students (Carver, 1992; Brysbaert, 2019), the estimated reading time for each text is between 5.0 and 5.6 minutes.

The texts are closely matched across several readability metrics (Appendix Table B1). Flesch-Kincaid Grade Levels fall between 15.1 and 15.9, corresponding to late undergraduate reading difficulty. Measures of lexical and syntactic complexity, including sentence length and word length, also suggest a high but consistent level of challenge across topics.

Table B1: Readability Metrics for Topic Texts

	Learning Topic:		
Metric	Blockchain	Carbon Capture	CRISPR
Flesch-Kincaid Grade Level	15.90	15.76	15.11
SMOG Index	16.75	17.28	16.50
Automated Readability Index	17.75	16.26	15.86
Type-Token Ratio	0.51	0.49	0.51
Average Sentence Length	21.59	24.15	20.80
Average Word Length	6.43	6.09	6.07
Word Count	1,253	1,333	1,399
Estimated Reading Time (250 wpm)	5.0 min	5.3 min	5.6 min

B.2 Writing Assessment Prompts

This appendix presents the analytical writing prompts developed for each of the three topics. For each topic, we designed two complementary prompts of comparable complexity, one for use in Session One and the other in Session Two. Each prompt requires participants to analyze relationships between concepts, evaluate the relative importance of different factors, and support their analysis with specific examples from the reading materials.

Blockchain Technology

1. "Examine how blockchain has grown beyond cryptocurrency. Analyze which two application areas have the strongest potential for transformative impact and explain the specific factors that support your analysis. Draw on examples from the reading to develop your response."

2. "Examine key differences between blockchain technology and traditional databases. Analyze which specific features make blockchain unique, which types of applications benefit most from these differences, and why. Support your analysis with concrete evidence or examples."

Carbon Capture

1. "Analyze the three main barriers to scaling carbon capture technologies (technical limitations, economic costs, and policy challenges). Identify which barrier you believe is most critical to overcome first, how it impacts the other challenges, and what approaches might address it. Support your analysis with specific examples."

2. "Compare direct air capture with point-source carbon capture approaches. Analyze the specific advantages and limitations of each method, identify which contexts each approach is better suited for, and explain what factors most influence their effectiveness. Draw on evidence from the reading to support your analysis."

CRISPR Gene Editing

1. "Analyze how CRISPR differs from previous gene editing technologies in terms of precision, accessibility, and versatility. Identify which two differences have been most consequential for scientific advancement, explain why these particular differences matter, and provide specific examples that demonstrate their impact."

2. "Examine the three main technical challenges that limit CRISPR today (delivery methods, off-target effects, and ethical considerations). Analyze which challenge is most important to address first, how it relates to the others, and what approaches show promise for overcoming it. Support your analysis with specific examples."

B.3 Human Essay Grading

We recruited independent graders through the Prolific platform to evaluate participants' analytical essays. Graders were blind to treatment condition, participant identity, and all other experimental variables. Each grader evaluated five essays and was compensated with a $15 fixed payment plus eligibility for a $10 lottery bonus. To incentivize consistency, graders were informed that their chances of winning the bonus were higher if their scores resembled those of other graders evaluating the same essays.

Before grading, each participant completed a comprehension check verifying their understanding of the task requirements and reviewed two example essays with suggested scores and detailed rationales. The examples were chosen to illustrate different quality levels—one strong response (receiving scores of 6–9 across dimensions) and one weaker response (receiving scores of 4–8)—so that graders could calibrate their standards. Graders were instructed to read each essay carefully, consult external resources to verify factual claims if needed, and maintain consistent scoring standards across all five essays.

Graders evaluated each essay on six dimensions using a 0–10 scale. *Accuracy of Content* measures the correctness and accuracy of the information presented in the essay. *Use of Evidence and Examples* assesses the integration of specific evidence, examples, and data from the reading to support arguments. *Relevance to the Prompt* evaluates how well the

essay addresses the specific prompt and stays on topic throughout. *Organization and Structure* measures the logical structure, coherence, flow, and effectiveness of transitions. *Writing Style and Clarity* assesses clarity, readability, grammar, and precision of language. *Overall Essay Quality* captures the grader's holistic assessment of the essay. Graders also reported their prior familiarity with each of the three topics (blockchain, carbon capture, and CRISPR) on a 0–10 scale before beginning the grading task.

Each essay was evaluated by three independent graders. We use the overall quality rating as our benchmark measure of essay quality and report results for all six dimensions separately.

B.4 AI Essay Grading

We complement our human grading with AI-generated scores to provide an independent assessment of essay quality. We use Anthropic's Claude Opus 4.6 model to grade each essay on the same six dimensions and 0–10 scale used by our Prolific graders. This section describes the grading procedure and the prompts used.

For each essay, we make six separate API calls—one per grading dimension—to avoid cross-contamination between scores. Each call includes a system prompt establishing the grading context, the dimension-specific rubric, two calibration essays with suggested scores, and the student essay to be graded. The calibration essays are the same examples shown to human graders during training, ensuring that human and AI graders share a common scoring anchor. The model returns a single integer score and a brief justification.

The system prompt is as follows:

```
You are an expert essay grader for a university research study. College students were
given 35 minutes to learn about a topic and write a 300-500 word analytical essay. You
must grade essays on a 0-10 scale following the rubric exactly. Be consistent and
calibrated using the example essays and scores provided.
```

For each dimension, the user prompt includes the dimension name, a description of what the dimension measures, and the scoring scale. The six dimensions and their scales are:

Accuracy of Content. Evaluate the factual correctness and depth of understanding demonstrated in the essay. Consider whether key concepts are accurately described, whether technical details are correct, and whether the student shows genuine comprehension of the topic.

0-3: Major factual errors or fundamental misunderstanding of the topic. 4-6: Some general knowledge but oversimplified or vague; key technical distinctions are glossed over. 7-8: Strong factual understanding; correctly describes basic mechanics, benefits, and drawbacks with some nuance. 9-10: Demonstrates advanced, detailed, and precise understanding of the topic with no factual errors.

Use of Evidence & Examples. Assess whether the essay draws on evidence from the provided reading materials. Consider how specific and well-integrated the references are.

0-3: No evidence or examples from the reading. 4-6: Minimal or vague use of evidence; general claims with only vague references like "the reading says..." without integrating specific details. 7-8: Good use of evidence with some specific examples from the reading; references are relevant but could be more detailed. 9-10: Excellent integration of specific evidence and examples from the reading throughout the essay.

Relevance to the Prompt. Measure whether the essay directly addresses the prompt requirements. Consider whether all parts of the prompt are addressed and whether the essay stays focused throughout.

0-3: Does not address the prompt or is largely off-topic. 4-6: Partially addresses the prompt but misses key aspects or lacks depth. 7-8: Mostly answers the prompt and touches on each required area; stays focused throughout. 9-10: Directly and thoroughly addresses every aspect of the prompt with analytical depth.

Organization and Structure. Evaluate the essay's organization, clarity of argument flow, and coherence. Consider whether there is a clear introduction, body, and conclusion, and whether transitions between ideas are smooth.

0-3: No discernible structure; ideas are disorganized or incoherent. 4-6: Basic structure with intro and conclusion, but loosely organized; paragraphs may be short and somewhat repetitive. 7-8: Clear and mostly effective structure; transitions could be smoother. 9-10: Excellent organization with smooth transitions, logical progression, and well-developed paragraphs.

Writing Style and Clarity. Assess the clarity, readability, grammar, and language precision of the essay. Consider vocabulary, sentence variety, and overall polish of the writing.

0-3: Very poor writing with major grammar issues or incoherent prose. 4-6: Very basic style; sentences are short and repetitive, vocabulary is limited. 7-8: Mostly clear and readable with minor awkward phrasing; lacks some polish or precision. 9-10: Excellent writing with varied sentence structure, precise vocabulary, and sophisticated prose.

Overall Essay Quality. Provide a holistic assessment of the essay's overall quality. Consider all previous dimensions together: accuracy, evidence, relevance, structure, and writing style.

0-3: Very poor quality; fails to demonstrate understanding or engage with the topic meaningfully. 4-6: Basic understanding demonstrated but lacks depth, sophistication, and engagement with the reading. 7-8: Solid response that answers the prompt with reasonable clarity and structure, though it may lack depth in evidence or style.
9-10: Exceptional quality; demonstrates deep understanding, excellent writing, strong evidence use, and thorough analysis.

The prompt also includes two calibration essays on carbon capture—the same examples used to train human graders—with suggested scores for the relevant dimension. After presenting the calibration material, the prompt provides the student's assigned topic, the essay prompt, and the essay text. The model is instructed to respond in a fixed format ("SCORE: [number]" followed by "JUSTIFICATION: [explanation]") to facilitate automated parsing. We set the maximum token length to 200 to keep responses concise.

Empty or very short essays (fewer than 20 characters) receive a score of zero on all dimensions without querying the model. For essays that the model fails to score after three retry attempts, we record the score as missing.

B.5 SAT Score Imputation

Middlebury College has test-optional admissions. Approximately 26 percent of participants (67 out of 260) did not take a standardized admissions test (SAT or ACT) and therefore have no test score on record. Because our double-lasso covariate selection procedure draws from a pool of potential controls that includes SAT score quintile bins, students without scores would all fall into a single "missing" bin, discarding potentially useful variation in academic preparation. To recover this variation, we impute SAT scores for missing students

using baseline covariates and assign them to quintile bins alongside students with observed scores.

We estimate an OLS regression of SAT scores on baseline characteristics among the 193 students with observed scores. The predictors are college GPA, age, gender, race and ethnicity indicators (white, Black, Latino, Asian), international student status, public and private high school indicators, academic field indicators (natural sciences, social sciences, humanities and arts), weekly study hours, and whether the student has declared a major. College GPA is the strongest predictor. We considered adding high school GPA to the prediction model, but it does not improve out-of-sample fit, likely because it is itself missing for a subset of students (particularly international students with non-comparable grading scales) and because college GPA already captures much of the same variation.

We assess predictive accuracy using leave-one-out cross-validation on the 193 students with observed SAT scores. The cross-validated R^2 is 0.215 and the root mean squared error is 125.7 points, relative to a sample mean of 1,388 and a standard deviation of 142. While the prediction model explains a modest share of the overall variance in SAT scores, the imputed values are sufficiently informative to place students into the correct region of the SAT distribution for the purpose of quintile bin assignment.

We predict SAT scores for all 67 students with missing values and use the combined observed and imputed scores to construct quintile bins via the `fastxtile` command. A missing indicator equal to one for imputed observations is included in the pool of potential controls, so the double-lasso procedure can account for any residual differences between observed and imputed groups. The imputed SAT scores are used only for bin assignment; they do not enter the regression directly as a continuous covariate. Our results are robust to alternative binning strategies (quintiles, deciles) and to excluding imputed observations entirely.

B.6 Classification of ChatGPT Conversations

During Session One, participants in the AI-allowed condition interacted with ChatGPT through monitored accounts, generating conversation logs that we subsequently classified using a large language model (Anthropic's Claude Opus). We describe the classification procedure here.

Prompt-Level Classification We classify each student prompt into one of six categories that mirror the self-reported usage types from Figure 3: (1) *Explaining concepts*—asking

ChatGPT to explain, clarify, or teach a concept from the reading; (2) *Summarizing readings*—asking ChatGPT to summarize or condense the reading material; (3) *Drafting responses*—asking ChatGPT to write, draft, or generate essay text; (4) *Proofreading*—asking ChatGPT to check grammar, spelling, or typos; (5) *Editing responses*—asking ChatGPT to revise, improve, rephrase, or restructure existing text; and (6) *Other*—anything that does not fit the above categories.

For each prompt, we provide the classifier with the full system prompt establishing the experimental context—that students were given a reading passage on a science topic and asked to write an essay—along with descriptions of each category. The classifier returns a single category label. Out of 435 prompts, 432 received clean single-label classifications; 3 returned verbose responses and were excluded. The resulting distribution is: Explaining concepts (48%), Other (22%), Drafting responses (12%), Editing responses (10%), Summarizing readings (6%), and Proofreading (2%).

To illustrate, Table B2 provides representative examples of student prompts classified under each category.

Table B2: Examples of Student Prompts by Classification Category

Category	Example prompt
Explaining concepts	"easy definition and understanding of crispr"
	"how might the cas9 attach to the wrong spot in crispr"
	"when is climate change irriversible"
Summarizing readings	"summary with bullet points and main takeaways from this document"
	"Give a brief overview on CRISPR technology and then explain why CRISPR differs from other gene editing technology"
Drafting responses	"wirte approximately 500 word response to the following prompt: analyze the three main barriers to scaling carbon capture technologies. . . "
	"Write a response paper of at least 500 words"
Editing responses	"Fix the grammar and make it flow smoother"
	"rewrite the whole thing now"
Proofreading	"how many words does it have"
Other	"keyboard shortcuts to copy and paste"
	"i think that policy challenges is most difficult what with trump in office and everything"

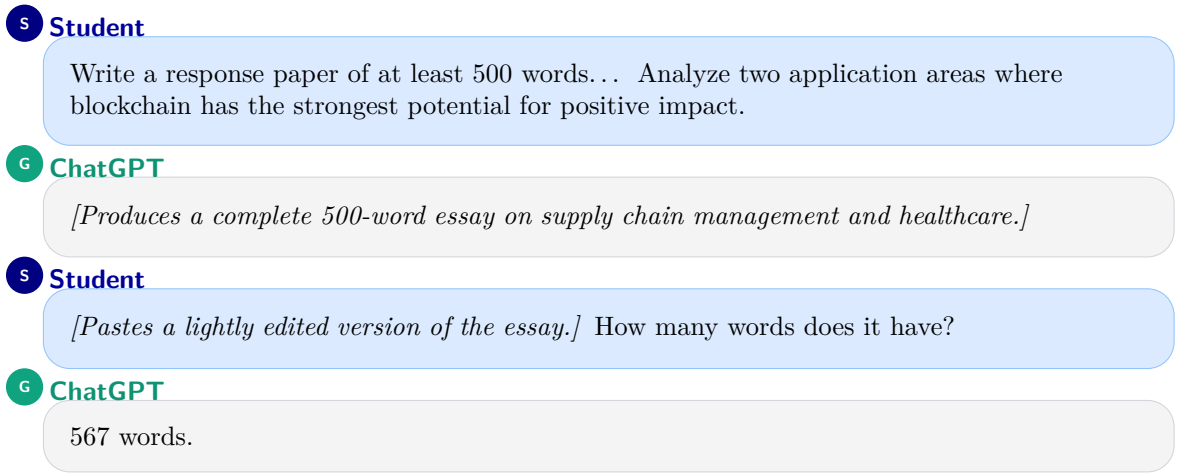
Notes: Each row shows a verbatim student prompt from the ChatGPT conversation logs. Prompts are reproduced exactly as typed, including spelling and grammatical errors.

Conversation-Level Classification We also classify each full conversation—the complete sequence of student and ChatGPT messages—into one of three categories: (1) *Automation*, where the AI does the work for the student (generating essay text, producing draft paragraphs, or creating content that could be pasted directly into the essay); (2) *Augmentation*, where the AI works with the student (explaining concepts, answering clarifying questions, or providing feedback on the student's own writing); and (3) *Mixed*, where the conversation contains substantial elements of both.

For each conversation, we provide the classifier with the full exchange between the student and ChatGPT, along with descriptions of each category. Out of 85 conversations, 82 received clean classifications; 3 returned verbose responses and were excluded. The resulting distribution is: Augmentation (55 percent), Automation (43 percent), and Mixed (2 percent).

To illustrate the distinction, we reproduce excerpts from conversations classified under each category. Appendix Figures B1 and B2 show selected exchanges; full transcripts are available via the hyperlinked URLs.

Figure B1: Automation Example: Student Delegates Writing to ChatGPT



Notes: Full transcript: https://chatgpt.com/share/681d3c21-9844-8012-b4f0-6a573222bd5b.

Figure B2: Augmentation Example: Student Uses ChatGPT as an Explainer

Student

Give me an in-depth explanation of blockchain technology up until the current moment.

ChatGPT

[Provides comprehensive overview: definition, origins, types of blockchains, major developments 2015–2025, the shift from Proof of Work to Proof of Stake.]

Student

What does proposing a new block mean? Also what does it mean that Satoshi Nakamoto proposed blockchain technology?

ChatGPT

[Explains both concepts in detail.]

Student

[Provides the assignment question.]

ChatGPT

[Suggests supply chain management and digital identity as two application areas, but does not write the essay.]

Notes: Full transcript: https://chatgpt.com/share/681d3d15-f188-8009-98d9-3d88b76fd1d5.

Use in Heterogeneity Analysis We use the conversation-level classification to construct student-level indicators of AI use type. A student is classified as an *automation user* if any of their conversations was labeled as Automation, and as an *augmentation user* otherwise. These indicators are defined only for treated students who used ChatGPT; control students are excluded from the classification by construction. In the heterogeneity analysis (Table 8), we compare each subgroup of treated students separately against the full control group.

B.7 Literature Comparison

We surveyed 20 experimental papers on generative AI and learning outcomes published or circulated between 2023 and 2025. We applied strict inclusion criteria to select comparable estimates for Figure 5. Appendix Table B3 summarizes the four included studies; below we describe the selection process and each study.

We required that each study satisfy five criteria: (1) a randomized experimental design; (2) an effect size reported in standard deviations or convertible from reported statistics; (3) a standard error or confidence interval (reported or back-calculable); (4) an outcome measuring learning on an unassisted assessment, excluding tasks where AI was available during evaluation; and (5) a clean no-AI control, rather than an active comparison such as human tutoring or structured hints. Criterion (4) deserves emphasis: several prominent studies report large effects on AI-assisted task performance, but these estimates confound learning with contemporaneous AI use and are not comparable to our design.

Table B3: Included Studies in Literature Comparison

Study	Domain	N	Design	Outcome	Estimates
Bastani et al. (2025)	Math	839	Cluster RCT	Unassisted exam	2
De Simone et al. (2025)	Language	759	RCT	Unassisted post-test	2
Lehmann et al. (2025)	Coding	176	Lab RCT (×2)	Unassisted post-test	2
Lira et al. (2025)	Writing	7,238	RCT (×3)	Retention test (24h)	3

Notes: N is the total sample across all arms included in our comparison. "Estimates" indicates the number of point estimates included in Figure 5. All studies compare AI access to a no-AI control on unassisted assessments.

Bastani et al. (2025) conducted a cluster-randomized trial with 839 Turkish high school students learning math on an online platform. Students were assigned to base GPT-4 access, GPT-4 with a tutoring prompt that withheld direct answers, or a no-AI control, and took an unassisted math exam after the practice period. Although base GPT-4 students solved more practice problems during the AI-assisted phase, this advantage reversed on the unassisted exam: the base group scored −0.19 SD relative to the control, while the tutor-prompted group scored −0.01 SD. We converted the authors' raw-scale standard errors to SD units by dividing by the control-arm standard deviation (0.277).

De Simone et al. (2025) randomized 759 primary school students in Nigeria to practice English language skills with an AI chatbot or a no-AI control. Students showed gains on unassisted post-tests in English comprehension (0.24 SD) and on a retention test two weeks later (0.21 SD). They also report a "total" estimate bundling AI and digital literacy skills

directly taught by the intervention; we exclude it because it captures treatment-specific content rather than general learning.

Lehmann et al. (2025) ran two laboratory experiments with Dutch university students learning Python programming (Study 2: $N = 107$; Study 3: $N = 69$). In both, treated students used ChatGPT during practice and then completed a 20-question unassisted coding post-test. Effects are 0.25 SD (Study 2) and 0.42 SD (Study 3), neither individually significant at conventional levels. Study 2 included an unintended copy-paste restriction that may have attenuated the treatment.

Lira et al. (2025) conducted three pre-registered experiments with large online samples ($N = 2{,}238$; 2,997; and 2,003). Participants used GPT-4-based coaching to learn professional cover letter writing and took an unassisted retention test 24 hours later. Effects ranged from 0.37 to 0.41 SD across the three studies, among the largest in our curated set of estimates from unassisted assessments.

We excluded the remaining 16 papers, most commonly for active control conditions (Kestin et al., 2025; Kreijkes et al., 2025; LearnLM Team, 2025), unreportable effect sizes, quasi-experimental designs without clean identification, or insufficient sample size. We document full exclusion reasons in the replication materials.

The random-effects grand mean in Figure 5 uses DerSimonian and Laird (1986) weights, accounting for between-study heterogeneity via a variance component estimated from study-level effects. The pooled estimate combines all 13 point estimates: 9 from the literature and 4 from our study.

C Student Beliefs About AI and Learning

This section analyzes student responses to an open-ended question included in the Session 2 survey: "In your opinion, how does generative AI (e.g., ChatGPT) affect student learning in college? Please explain your reasoning." Of 210 students who completed Session 1, 204 (79.7 percent) provided a non-missing response. We analyze these responses in two ways: sentiment analysis to validate the measure, and LLM-based classification to characterize the content thematically.

C.1 Validating the Open-Ended Response Measure

We begin by validating our open-ended response measure using VADER (Valence Aware Dictionary and sEntiment Reasoner), a lexicon-based sentiment analysis tool (Hutto and Gilbert, 2014). VADER assigns each response a compound sentiment score ranging from -1 (most negative) to $+1$ (most positive). We then examine whether sentiment scores correlate with AI usage, both observed and self-reported.

Appendix Figure C1 presents binned scatterplots relating compound sentiment to two measures of AI use. Panel A restricts to AI-allowed students and shows the relationship with whether a student actually used the provided ChatGPT account during Session 1 (observed from server logs). The coefficient is small and not statistically significant ($\hat{\beta} = 0.090$, SE $= 0.082$), indicating that sentiment does not predict experimental take-up. This is reassuring: beliefs about AI's effect on learning do not appear to drive compliance with the experimental treatment.

The picture changes when we consider general AI usage patterns. Panel B reveals a strong and significant relationship with the frequency of AI use for academic work during the semester ($\hat{\beta} = 0.430$, SE $= 0.162$, $p = 0.009$). Students who express more positive views about AI's effect on learning report using AI tools more frequently in their regular coursework. This mirrors the pattern documented in our companion survey paper (Contractor and Reyes, 2025), where sentiment toward AI similarly predicts general adoption but not experimental behavior. The contrast between experimental and general usage suggests that the decision to use AI when it is provided in a controlled setting operates through different channels than the decision to adopt AI in everyday academic life.

C.2 Thematic Classification of Responses

To analyze the content of responses systematically, we classified each one using a large language model (Claude Opus). We first developed a codebook inductively by providing all 204 responses to the model and asking it to derive categories that balance parsimony against meaningful distinctions. After reviewing and refining the resulting categories, we used the same model to classify each response individually, assigning both a single primary label and a set of multi-label codes. The ten categories, with their frequencies, are reported in Appendix Table C1.

The dominant theme is conditionality. Half of all responses (50.0 percent) frame AI as a "double-edged sword" whose impact depends on how students use it. As one student put it: "Generative AI can be a powerful tool for students to use, streamlining the process of researching a topic by finding sources and information in seconds instead of hours. [...] However, it can also be used as a way to more easily be academically dishonest. [...] In this sense it is a double-edged sword, with each individual student's use determining whether it is helpful or hurtful for their education." This conditional framing is consistent with the observation that most students have already adopted AI and are navigating the boundary between productive and counterproductive uses.

Among directional responses, concerns about learning dominate. The most common non-conditional category is harm to critical thinking and deep learning (13.7 percent), where students worry that AI undermines the cognitive effort that produces genuine understanding. One student noted: "I think that generative AI negatively impacts a student's ability to critically think about a topic and construct an argument about that topic." Dependency and over-reliance (7.4 percent) captures a related but distinct concern about behavioral patterns. As one student described: "It becomes a crutch. Things I used to use my own brain for, I now rely on ChatGPT to do. It's like when you start to rely on a calculator—you begin using it for simple calculations that you could do in your head but become lazy."

Positive views emphasize AI as a learning aid (10.8 percent) and efficiency tool (9.3 percent). Students in the former category describe using AI to understand concepts, break down jargon, and generate practice questions. Those in the latter focus on time savings: "It allows us to allocate our time to more important things other than 500 pages of reading a night." A small number of responses raise concerns about academic dishonesty (2.5 percent), information quality (1.5 percent), knowledge retention (1.5 percent), or equity and future readiness (1.5 percent).

When we allow multi-label coding—so that a single response can be tagged with multiple themes—the picture changes. Under this approach, 57.8 percent of responses mention AI as a learning aid, 53.4 percent mention efficiency benefits, 42.2 percent raise concerns about critical thinking, 35.8 percent mention academic dishonesty, and 29.9 percent discuss dependency. The gap between single-label and multi-label frequencies confirms that most responses touch on several themes, with the conditional framing serving as the primary organizational structure within which students articulate specific benefits and concerns.

C.3 Do Beliefs Predict AI Usage?

We next examine whether thematic categories predict AI usage. Using multi-label indicators for each category, we regress three measures of AI use on category dummies, controlling for strata fixed effects and clustering standard errors at the student level (Appendix Table C2).

Two patterns stand out. First, beliefs about AI as a learning aid (C01) strongly predict higher self-reported frequency of AI use ($\hat{\beta} = 0.811$, $p < 0.001$), while concerns about critical thinking harm (C03) predict lower frequency ($\hat{\beta} = -0.704$, $p < 0.001$). These coefficients are large in magnitude: mentioning AI as a learning aid is associated with using AI about 0.8 points higher on a 1–5 frequency scale (from "Never" to "Very frequently"), while expressing concerns about critical thinking is associated with a 0.7-point decrease. Second, mentioning equity, accessibility, and future readiness (C09)—a small but distinctive group—is associated with significantly more prompts sent to the provided ChatGPT account ($\hat{\beta} = 1.801$, $p = 0.030$). Students who view AI as a necessary professional skill tend to use it more.

Notably, mentioning academic dishonesty concerns (C04) predicts *lower* experimental ChatGPT use ($\hat{\beta} = -0.123$, $p = 0.085$). Students who worry about AI enabling cheating are less likely to use the provided ChatGPT account during the experiment, suggesting that integrity concerns are not merely performative—they correspond to behavioral restraint even when AI access is explicitly authorized.

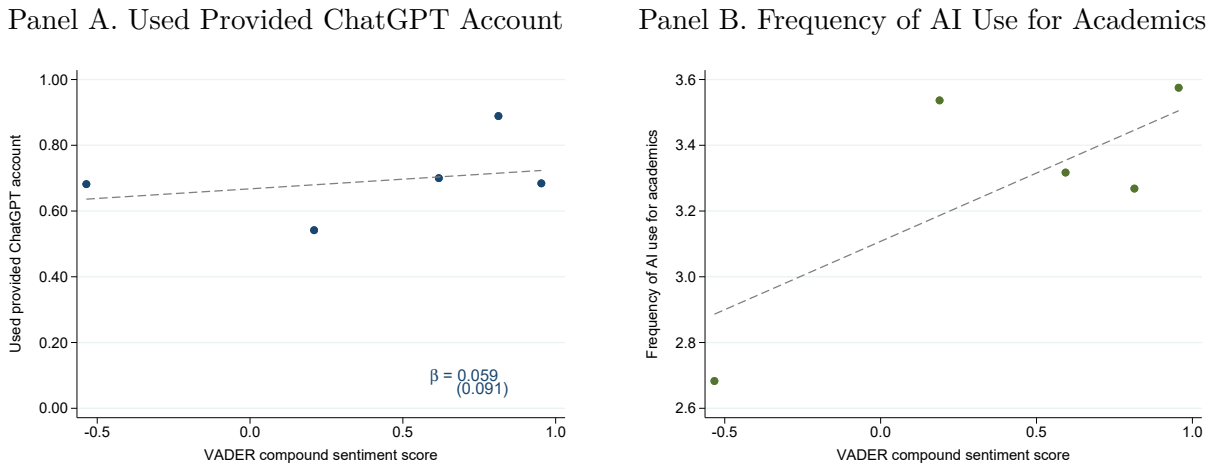
C.4 Discussion

Two themes emerge from this analysis. First, the overwhelming prevalence of conditional framing (50 percent of primary labels, 61 percent in multi-label coding) suggests that students have internalized a nuanced view of AI as a technology whose impact depends

on usage. This is not hedging or indifference; students who frame AI conditionally still articulate specific benefits and harms, but they resist assigning a single directional verdict. This finding is consistent with the survey paper's observation that students draw sharp distinctions between AI uses that enhance learning and those that substitute for it.

Second, beliefs predict behavior in intuitive ways. Students who see AI as a learning aid use it more frequently; those who worry about critical thinking use it less. These associations hold conditional on strata fixed effects, suggesting they are not merely artifacts of course-level differences in AI exposure. The disconnect between beliefs and experimental take-up—where sentiment and thematic content are largely unrelated to whether students used the provided ChatGPT account—reinforces the interpretation that experimental compliance is driven by immediate task characteristics rather than general attitudes toward AI.

Figure C1: Relationship Between AI Sentiment and AI Usage



Notes: This figure presents the relationship between AI sentiment and AI usage. Panel A restricts to students in the AI-allowed condition and shows the proportion who used the provided ChatGPT account during Session 1 (from server logs). Panel B shows the self-reported frequency of AI use for academic purposes during the semester (1 = Never, 5 = Very frequently) for all respondents. Each point represents the mean outcome for respondents within sentiment score quintile bins. Sentiment scores are compound scores computed using Hutto and Gilbert (2014)'s VADER algorithm applied to responses to the open-ended question about AI and learning. Positive values indicate positive sentiment and negative values indicate negative sentiment. The dashed lines show OLS best-fit lines estimated on the microdata, with coefficients and standard errors (in parentheses) displayed in each panel.

Table C1: Thematic Classification of Open-Ended Responses

Category	N	% (single)	% (multi)
Learning aid / conceptual understanding	22	10.8	57.8
Efficiency / time-saving tool	19	9.3	53.4
Harm to critical thinking and deep learning	28	13.7	42.2
Academic dishonesty and shortcutting	5	2.5	35.8
Dependency / over-reliance	15	7.4	29.9
Conditional / depends on usage	102	50.0	61.3
Information quality / accuracy concerns	3	1.5	21.1
Knowledge retention / memory impairment	3	1.5	5.9
Equity, accessibility, and future readiness	3	1.5	14.2
No clear opinion / non-substantive	4	2.0	2.0
Total	204	100.0	

Notes: This table reports the distribution of thematic categories from LLM-based classification (Claude Opus) of 204 open-ended responses to the question: "In your opinion, how does generative AI (e.g., ChatGPT) affect student learning in college? Please explain your reasoning." The "% (single)" column shows the share of responses assigned each category as the single best label. The "% (multi)" column shows the share of responses that mention each theme under multi-label coding, where a single response can receive multiple category tags. Categories and definitions were derived inductively from the full set of responses before classification.

Table C2: Thematic Categories Predicting AI Usage

	Used ChatGPT (from logs) (1)	AI frequency (1–5 scale) (2)	Nr. prompts sent (3)
Learning aid	0.020	0.811***	0.766
	(0.068)	(0.169)	(0.467)
Efficiency tool	−0.016	0.435**	−0.168
	(0.069)	(0.174)	(0.566)
Critical thinking harm	−0.071	−0.704***	−0.042
	(0.068)	(0.174)	(0.536)
Academic dishonesty	−0.123*	0.107	−0.619
	(0.071)	(0.184)	(0.482)
Dependency	−0.005	−0.116	0.396
	(0.076)	(0.192)	(0.626)
Accuracy concerns	0.133	−0.282	0.658
	(0.087)	(0.216)	(0.721)
Retention harm	−0.087	−0.102	0.129
	(0.129)	(0.451)	(1.003)
Equity / readiness	0.173*	0.176	1.798**
	(0.105)	(0.278)	(0.820)
N	204	204	204

Notes: Each cell reports the coefficient from a separate bivariate regression of an AI usage measure on the indicated multi-label category indicator, with strata fixed effects. Column (1): indicator for having used the provided ChatGPT account during Session 1 (from server logs). Column (2): self-reported frequency of AI use for academic purposes (1 = Never, 5 = Very frequently). Column (3): number of prompts sent to the provided ChatGPT account. Standard errors clustered at the student level in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.