NET WORK LISER

Discussion Paper Series

IZA DP No. 18792 July 2026

Experimental Evidence on the Learning Impact of Generative AI

Zara Contractor
Middlebury College

Germán Reyes
Middlebury College
and IZA@LISER

IZA DP No. 18792 July 2026

Experimental Evidence on the Learning Impact of Generative AI*

Abstract

We study how generative AI affects student learning in a randomized experiment. In proctored, in-person sessions, undergraduates learn about an unfamiliar topic and write an analytical essay with or without access to off-the-shelf generative AI, then complete unaided assessments immediately and one week later. We measure learning with knowledge tests (factual and conceptual understanding) and open-ended essays (higher-order skills). AI access raises immediate test scores by 0.27 standard deviations. These gains persist one week later. Essay quality, by contrast, changes little while students have AI access but improves in style and relevance one week later, when students write unaided. These delayed gains are larger among *augmentation users*—who use AI to explain concepts rather than generate text—whereas *automation users'* short-run quality gains vanish once AI is removed. We find evidence for two mechanisms behind the learning gains: students shift time away from drafting text and toward reading and searching for information, and they report greater learning enjoyment.

JEL classification

I21, I23, J24, O33, C93, D83

Keywords

generative AI, human capital, learning, higher education, randomized experiment

Corresponding author

Germán Reyes
greyes@middlebury.edu

* For helpful discussions and comments, we thank Chris Campos, Jeff Carpenter, Amy Collier, Sam Hirshman, David Huffman, Guy Ishai, Cory Koedel, Seunghoon Lee, Peter Matthews, David Munro, Caitlin Myers, Ted O'Donoghue, Evan Riehl, Andrea Robbett, Nick Swanson, and participants at the Liberal Arts Colleges Behavioral and Experimental Economics Workshop (LACBEE), the NBER Economics of Education Spring Meeting, the New England Experimental Economics Workshop, the Northeastern Economics of Education Workshop, the Royal Economic Society Conference, the Society of Labor Economists Annual Meeting, and seminars at Cornell University, Middlebury College (Carol Rifelj Faculty Lecture), Ohio University, the University of Missouri, and the University of Texas Rio Grande Valley. Financial support from 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 study was reviewed and approved by the Institutional Review Board of Middlebury College. Survey instruments and additional materials are available in the Supplementary Materials.

1 Introduction

Generative artificial intelligence (AI) is now widely used in higher education. ChatGPT has over 900 million weekly users, college-age users send nearly half of its messages, and more than a quarter of those messages relate to learning (Chatterji et al., 2025; OpenAI, 2025). Despite this widespread use, evidence on whether AI helps students learn is scarce. Much of the available evidence comes from workplace settings, in which a growing literature documents large productivity gains from AI access (e.g., Noy and Zhang, 2023; Brynjolfsson et al., 2025). However, these studies measure job performance and not human-capital accumulation. AI tools could enhance learning by acting as on-demand tutors that explain difficult concepts and provide personalized feedback (Nickow et al., 2024), or undermine learning by offloading the cognitive effort that learning requires (Risko and Gilbert, 2016). Whether generative AI builds or erodes human capital remains an open empirical question.

We study whether off-the-shelf AI affects immediate and longer-term learning in a common class of academic tasks: time-bounded assignments in which students must learn unfamiliar material, identify relevant information, and demonstrate understanding through writing and tests. Our randomized experiment with 211 undergraduates unfolds over two sessions approximately one week apart. In the first session, students learn about a technical topic and write an analytical essay. We randomly assign students to an *AI-allowed* or an *AI-forbidden* condition. The AI-allowed group may use any generative AI tool along with traditional resources, while the AI-forbidden group may use only traditional resources. We monitor AI usage and compliance through direct proctor observation, ChatGPT conversation logs, and self-reports. In the second session, all students complete another test and write an essay on the same topic, this time without AI or any other resources.

We measure learning along two dimensions. Knowledge tests target factual and conceptual understanding: we ask students to recall key information and identify core mechanisms. Open-ended essays target higher-order skills: we ask students to analyze relationships between concepts and support their arguments with specific evidence. Human and AI graders evaluate each essay along several dimensions, such as writing style and accuracy of content. Beyond these subjective assessments, we also measure objective linguistic features of the essays, such as length, readability, lexical diversity, textual similarity among student essays, and the fraction of text flagged as AI-generated by a state-of-the-art AI detector.

We first document that our random assignment generated substantial variation in AI usage. About 68 percent of students in the AI-allowed condition used generative AI during

the first session, compared with near-zero usage in the control group. Users value this access: 88 percent report that AI is helpful for the learning task. ChatGPT conversation logs reveal that students use AI most often to explain concepts (57 percent of treated students), followed by drafting responses (31 percent) and summarizing readings (17 percent).

We turn next to our main result: AI access improves learning outcomes. Students with AI access score 6.7 percentage points (pp) higher on the immediate knowledge test on a baseline of 56.3 percent. Standardized by the control-group standard deviation (SD), this is equivalent to a 0.27 SD gain in test scores. Point estimates are largest in the middle of the performance distribution, with smaller effects at the tails. Turning from where in the performance distribution gains land to which students gain, gains tend to be larger in the upper ability quartiles (by GPA and SAT) than at the bottom, suggesting that AI access may widen learning gaps. The gains persist approximately one week later: AI-allowed students score 5.1 pp higher on the retention test. Thus, about 76 percent of the immediate effect persists.

AI access also strengthens the higher-order skills that essays measure, but the gain surfaces only after AI is removed. In Session One, a higher share of treated students' text is flagged as AI-generated and their essays trend longer, with shorter sentences and simpler vocabulary. Their essays are no more homogeneous than control students', contrary to the convergence documented in workplace settings (Brynjolfsson et al., 2025). Overall quality rises only slightly: averaging human and AI graders, the gain is small and imprecise. Because these essays were partly generated by AI, these differences blend AI-produced text with any learning. To separate the two, we turn to the Session Two essays, written one week later without AI. There, the AI-detection and stylistic effects fade, while quality gains emerge: writing style and clarity and relevance to the prompt both rise significantly. These quality gains are consistent with the persistent test-score effects.

Treatment effects—and their persistence—vary substantially with how students use AI. Using treated students' ChatGPT conversation patterns, we classify 49 percent of AI users as "augmentation" users (who work *with* AI) and 32 percent as "automation" users (who use AI to do the work *for* them). This distinction shows up in both usage and output: augmentation users prompt AI mainly as a personalized tutor, while automation users prompt it to draft their essays for them and have substantially higher AI-detection rates in their submissions. Augmentation users' Session One test-score gains persist one week after AI is removed, and their unaided Session Two essays trend higher in quality. Automation users' Session One essay-quality gains, by contrast, fade entirely.

To investigate mechanisms, we examine the inputs to learning. AI access does not change total learning time but tilts its mix away from producing text: treated students spend 12 percent less time on writing activities (consistent with AI doing some of the writing for them) and 10 percent more time reading and searching for information. AI access also makes learning more enjoyable: treated students report 13 percent higher enjoyment ratings. We also find that AI access raises rule violations on the first test, though these violations account for at most a modest share of the test-score gains.

We also study students' beliefs about AI's effect on their learning by asking them to estimate how much AI access changed (or would have changed) their test scores. Both treated and control students expect positive effects, but only the treated group's prediction comes close to our estimate; the control group overestimates it. Across subgroups, perceived gains track actual gains: the subgroups that benefit most perceive the largest gains. To understand why students expect these effects, we elicit an open-ended account of how AI affects student learning in general and, following Andre et al. (2026), code each response as a causal graph from AI use to learning. Most students name both a channel through which AI helps learning and one through which it harms it, condition the effect on how AI is used, and spontaneously organize the channels around the augmentation-versus-automation distinction that moderates our treatment effects. Students, in short, grasp the mechanisms from the outset but only gauge the magnitudes accurately after using AI themselves.

Our results contribute to an emerging literature on the learning impacts of AI. We contextualize our estimates through a meta-analysis in the main text, but note two challenges for interpreting existing evidence. First, many studies evaluate AI systems that are scaffolded, restricted, or otherwise customized beyond a standard off-the-shelf chatbot.[1] These designs inform the particular system studied but speak less directly to the chatbots students use in practice. Second, some studies compare AI with active alternatives rather than a business-as-usual counterfactual, so the estimates reflect AI's value relative to those alternatives rather than its absolute effect on learning.[2] Our design addresses

[1]For example, Bastani et al. (2025) test GPT-4 with guardrails that provide teacher-designed hints rather than direct solutions in Turkish high-school math classes; Xu et al. (2025) add metacognitive prompts to a generative-AI learning environment for Chinese college students; Kim et al. (2025) restrict AI to tutoring after students attempt each problem on a math practice platform; De Simone et al. (2025) study teacher-guided sessions with Microsoft Copilot in after-school English classes in Nigeria; Poulidis et al. (2025) compare a system-regulated AI that auto-delivers tips at key moments with on-demand AI in a 12-week chess training program; and Kumar et al. (2025) expose online adults practicing math problems to explanations pre-generated by GPT-4 under a hidden tutoring prompt, rather than to the chatbot itself.

[2]For example, Kestin et al. (2025) compare AI with in-class active learning in a college physics course,

both challenges: treated students access off-the-shelf AI tools, while control students learn the same material without any AI access. Beyond this design, we contribute by measuring knowledge persistence; distinguishing augmentation from automation uses; tracing how AI shifts the inputs to learning; assessing multiple learning dimensions rather than test scores alone; and comparing students' beliefs about AI's effects with those we estimate.

We also contribute to the literature on digital technologies and student learning. Prior research has examined laptops and personal computers (Malamud and Pop-Eleches, 2011; Carter et al., 2017; 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). Expanding access to general-purpose technologies tends to yield limited, mixed, or negative effects on academic performance. Computer-assisted learning systems—which tailor content to students' levels and provide immediate feedback—produce larger and more consistent gains (see Bulman and Fairlie, 2016; Escueta et al., 2020, for reviews). Generative AI offers these capabilities: it can generate personalized explanations and respond flexibly to students' questions in real time, but it can also do the work for students. Whether these capabilities translate into learning gains depends on how students use AI—actively constructing understanding builds more durable knowledge than passively accepting AI output (Chi and Wylie, 2014)—as our augmentation-versus-automation results show.

Finally, we contribute to the literature on the productivity effects of AI. Prior studies document substantial productivity gains from AI access in professional writing, software development, consulting, customer support, legal work, team-based problem solving, radiology, and taxi driving (Noy and Zhang, 2023; Peng et al., 2023; Choi et al., 2024; Jia et al., 2024; Brynjolfsson et al., 2025; Dell'Acqua et al., 2025; Kanazawa et al., 2025; Baird et al., 2026; Cruces et al., 2026; Cui et al., 2026; Dell'Acqua et al., 2026; Goldsmith-Pinkham et al., 2026). This literature measures task performance: how well workers complete tasks when given access to AI. Yet productivity gains can also accrue through skill accumulation—how effectively individuals build durable human capital—but the few studies that speak directly to learning in work settings provide mixed results. Brynjolfsson et al. (2025) show that AI facilitates worker learning among customer-support agents, while Budzyń et al. (2025) document patterns consistent with deskilling among endoscopists after

LearnLM Team (2024) with static pre-written hints and human tutoring across a range of tutoring scenarios, Kreijkes et al. (2026) with note-taking on reading-comprehension tasks in secondary schools, and Chung et al. (2026) compare an AI tutor that adapts the order and difficulty of practice problems to each student with a fixed-sequence version of the same tutor in a high-school Python course.

routine AI exposure, and Shen and Tamkin (2026) find that software engineers who learned an unfamiliar programming library with an AI assistant scored substantially worse on a subsequent unassisted test. We provide experimental evidence that AI access can improve skill accumulation in an academic context, and thereby identify a channel through which AI may raise long-run productivity beyond its immediate effects on task performance.

2 Experimental Design

The experiment consists of two in-person sessions conducted approximately one week apart. Survey instruments are available in the Supplementary Materials.

2.1 Session One

Overview. The first session employs a between-subjects design with random assignment to either an AI-allowed or AI-forbidden condition. We schedule eight time slots with two parallel labs per slot, one for each treatment arm. For each time slot, we randomly assign one lab to the AI-allowed condition and the other to the AI-forbidden condition. We randomly assign each student who signed up for a time slot to a specific computer in one of these two labs. Each workstation has a privacy divider to minimize distractions and prevent students from viewing other screens (see Appendix Figure A1). We administer all components of the session through Qualtrics, a survey platform that tracks each student's progress, enforces section time limits, and records how long students spend on each component.

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 and asking students to put their phones away for the duration of the session. We tell students that their primary task involves learning about a prespecified topic and that they will demonstrate their understanding through a series of assessments. To ensure comprehension of the experimental procedures, students complete a series of comprehension checks, with laboratory staff providing clarification if needed. Students correctly answer an average of 4.77 out of 5 comprehension questions, and the median student answers all questions correctly.

Topic Assignment and Baseline Assessment. We select three topics rich in factual content, about which most students have limited prior knowledge: blockchain technology, carbon capture systems, and CRISPR gene editing. Students are randomly assigned to one of these three topics and cannot change their assignment. Students then complete a five-question multiple-choice test to assess their baseline knowledge of the assigned topic. In all tests, the order of questions and answer options within each question is randomized. Students may not use notes, the internet, or AI tools during tests.

Learning Phase. Following the baseline test, students 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, students also write an approximately 500-word analytical essay demonstrating their understanding of the topic. Each essay prompt asks students to apply critical thinking skills by analyzing relationships between concepts, comparing elements, and supporting their analysis with specific evidence (see Appendix B.2).[3] For each topic, we develop two prompts; each student is randomly assigned one for Session One and the other for Session Two. We provide all students with a link to an introductory reading about their assigned topic, created to be of comparable length and reading difficulty across the three topics.[4]

Students can allocate their time freely across reading, searching, drafting, editing, or other activities such as surfing the web. 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, students complete a brief survey that measures self-assessed current knowledge of the topic (0–10 scale); time allocation 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. Students next complete five multiple-choice questions to test factual knowledge and conceptual understanding of the assigned topic. These questions differ

[3]For example, a student 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 their analysis with specific examples.

[4]Texts range from 1,253 to 1,399 words (Appendix Table B1). At an average silent reading speed of approximately 250 words per minute for college students (Carver, 1992; Brysbaert, 2019), the estimated reading time is 5.0 to 5.6 minutes per text. The full text of the learning materials is available in the Supplementary Materials.

from the baseline test. Students complete this assessment without access to external resources, the internet, or AI tools regardless of treatment condition.

Waiting Room. Students who complete all tasks before the 60-minute session ends 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. We tell students that we record their responses only after they click this button. This prevents early departures from disrupting others or creating social pressure effects. During this waiting period, students may browse the internet freely.

Treatment. We randomly vary students' access to generative AI tools during the learning phase. The treatment has two components: explicit instructions about whether AI use is permitted, and a logged-in ChatGPT account for students in the AI-allowed condition.

Students in the *AI-allowed* condition receive explicit instructions that permit the use of any generative AI tools during the learning phase. The instructions are delivered orally through the proctor-read welcome script and in writing through the survey interface (Appendix Figure A2, Panel A). 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 on one of the tabs, so you don't need to use your personal account." To ease access, each computer has three browser tabs already open: a dedicated ChatGPT (GPT-4o) account, Wikipedia, and the Middlebury College Library website.

Students in the *AI-forbidden* condition receive explicit instructions prohibiting use of generative AI tools. The instructions are delivered orally through the proctor-read welcome script and in writing through the survey interface (Appendix Figure A2, Panel B). The script states: "You are not allowed to use generative AI tools such as ChatGPT, Claude, or other AI assistants." Each computer has 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 students 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 AI-forbidden 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, the platform periodically captures students' writing interfaces to detect sudden appearances of large text blocks that may indicate copying from external sources.

Compensation and Incentives. Students receive both attendance compensation and performance-based incentives. For attendance, students receive $5 for completing Session One alone but $50 for completing both sessions. For performance, students earn lottery tickets: one ticket per correct response on each test and one per point on their rounded quality score. At the end of the experiment, we conduct a drawing in which 30 lottery tickets are randomly selected, with each winning ticket worth $100. To further support attendance, we send email and text reminders on the day of each scheduled session and let students schedule the second session flexibly, rather than requiring exactly seven days between sessions.

2.2 Session Two

Session Two takes place approximately one week after Session One.[5] This session measures knowledge retention when all students—regardless of Session One treatment 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 that uses the complementary prompt developed for each topic.

Session Two concludes with a questionnaire on four topics. First, beliefs about AI's impact on their own and others' test scores. Second, self-reported AI skills (Likert scale), frequency of use across academic tasks, and when students began using AI. Third, retrospective Session One behavior, including which tools they used and for what purposes (explaining concepts, summarizing, drafting, proofreading, editing) and whether they studied the topic between sessions. Fourth, an open-ended question about generative AI's impact on learning.[6]

[5]When scheduling, we encouraged students 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 6.99 days. Most students scheduled their second session exactly seven days after Session One (61.4 percent), and 80.5 percent attended within six to eight days.

[6]We note four main deviations from the experimental protocol that occurred during early time slots.

3 Data and Research Design

3.1 Sample and Recruitment

The experiment took place at Middlebury College during Spring 2025. At the time of the experiment, AI adoption at Middlebury exceeded 80 percent (Contractor and Reyes, 2026), so our treatment varies AI access among students already familiar with the tool. We recruited students through campus-wide email announcements, posted flyers, student group chats, and a dedicated recruitment website. To minimize selection concerns, we framed the study as "a study on learning" without mentioning AI in any recruitment materials.

3.2 Summary Statistics and Balance

Table 1 reports summary statistics for three groups: the 256 students who completed the sign-up survey (column 1), the 211 (82.4 percent) who attended Session One (column 2), and the 204 (79.7 percent) who attended both sessions (column 3). We focus on column 2. The typical student is 20.1 years old, 35.1 percent are male, 49.8 percent white, 26.5 percent international, and just over half (54.0 percent) attended a public high school (Panel A). The mean GPA is 3.68 and the mean SAT score is 1386 (Panel B).[7] The majority of students (96.7 percent) returned for Session Two (Panel C). Students self-reported little baseline knowledge of their assigned topic (1.5 on a 0–10 scale) and, consistent with this self-assessment, answered only 30.3 percent of baseline multiple-choice questions correctly (Panel D).

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

First, the initial Session One verbal welcome script did not mention AI permissions or restrictions; students 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 students 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 students from completing the essay on the platform. We added extra time, asked affected students to write their essays in Microsoft Word, and entered those essays into the dataset manually. All analyses include session fixed effects to account for these protocol variations.

[7] We asked students 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.

10 percent level. An F-test does not reject joint balance ($F = 1.34$, $p = 0.173$). Measures of academic ability—strong predictors of learning outcomes—are well balanced: mean GPA (3.69 versus 3.67) and SAT scores (1392 versus 1383) 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 82.3 percent) and no differential attrition between sessions.

3.3 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 student i, AI-Allowed$_i$ is an indicator that equals one if the student 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 fixed effects for the randomization strata (the eight Session One time slots) and additional controls selected through the double-lasso procedure (Belloni et al., 2014), which selects controls that predict either the outcome or the treatment assignment from a pool of potential covariates.[8] We report heteroskedasticity-robust standard errors.

The coefficient of interest, β, measures the causal effect of AI access on a given outcome. 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 AI. This may be the relevant parameter for policy evaluation: institutions can choose whether to permit AI use, but they cannot force students to use it. To estimate the effect of actual AI use rather than assignment, we complement our ITT estimates with a treatment-on-the-treated (TOT) analysis. We instrument actual AI use with the randomly assigned treatment status (AI-allowed) and estimate by two-stage least squares (2SLS). These estimates identify the

[8]The pool of potential controls includes: (1) demographic characteristics—age, gender, race/ethnicity indicators (Black, Latino, Asian, white), international student status, and public/private high school attendance; (2) academic background—cohort fixed effects, declared major status, field-of-study indicators (natural sciences, social sciences, humanities/arts), self-reported study hours per week, college GPA, SAT score, and high school GPA (each entered continuously and as decile bins), and an indicator for taking a standardized admissions test; (3) baseline measures—pre-learning self-assessed knowledge, pre-learning test score, prior experience with similar tasks, and comprehension check score; (4) experimental design features—topic fixed effects, writing prompt version, Session Two assessment order, and lab location; and (5) missing-data indicators. For 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.

local average treatment effect (LATE) for compliers—students whose treatment assignment influenced their AI use.

3.4 Main Outcomes

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

AI Usage. As a first-stage measure, we examine whether students used generative AI during the learning phase. Our primary measure is an indicator that equals one if the ChatGPT account provided to the student during Session One contained at least one prompt. We measure intensity through the number of prompts sent and the number of distinct conversations (i.e., separate chat threads) initiated. We complement these with self-reported measures of whether students used any generative AI tool, ChatGPT specifically, or alternative AI models (Claude, Gemini, Copilot, DeepSeek, or Llama).

Test-Based Learning Outcomes. We measure knowledge acquisition and retention using self-reported and objective measures. Self-assessed knowledge captures students' subjective evaluation of their understanding on a 0–10 scale, with 0 indicating "I know nothing about this topic" and 10 indicating "I am an expert."[9] The fraction correct is the proportion of multiple-choice questions answered correctly (out of five questions in Session One, and 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 (separately for each session). We also define binary indicators for scoring above multiple thresholds to trace treatment effects across the score distribution.

Essay-Based Learning Outcomes. Essays measure higher-order skills such as analytical reasoning, synthesis, and argumentation. We recruited 311 graders through Prolific, restricting eligibility to individuals holding a master's or PhD degree. Each essay was independently evaluated on a 0–10 scale by three or four graders blind to treatment condition and student identity; each grader scored five essays. Graders provided an overall quality rating and scores on five dimensions: accuracy of content, use of evidence and examples, relevance to the prompt, organization and structure, and writing style and clarity (see Appendix B.3 for details and Supplementary Materials for the grading instrument). We

[9]This measure may capture dimensions of learning that 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.

also constructed a quality index as the average of the five dimension scores, each on the 0–10 scale; we standardize it by the control-group standard deviation when we report it in SD units. We complement human grading with AI grading: a large language model (LLM) evaluates each essay on the same dimensions and rubric (see Appendix B.4 for the procedure and prompts).[10] For our main regressions, we use the average of human and AI grader scores for each dimension and show robustness to using only the human or only the AI ratings.

Objective Linguistic Features. In addition to subjective grader ratings, we compute three objective dimensions of students' writing style: length, readability, and lexical diversity. We measure length through token, word, and sentence counts. We measure readability through sentence length, syllables per word, the Flesch-Kincaid grade level, and the Flesch Reading Ease score. We measure lexical diversity 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 SD 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-to-read text), and a lexical diversity index.

To examine whether AI access homogenizes student writing, we follow Brynjolfsson et al. (2025) in computing sentence embeddings. 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 reading material similarity as the cosine similarity between the essay and the provided reading material.[11] Finally, we use Pangram, a state-of-the-art AI content detector (Emi and Spero, 2024; Masrour et al., 2025; Thai et al., 2026), to measure the fraction of text classified as AI-generated.[12] We also use Pangram to measure the fraction flagged as plagiarized.[13]

[10] Two considerations motivate our use of AI grading: a growing literature validates LLMs as evaluators of text quality (Chiang and Lee, 2023; Liu et al., 2023; Zheng et al., 2023), and our learning topics (blockchain, carbon capture, CRISPR) are technical subjects on which human graders recruited from a general population may have limited expertise.

[11] As a validation exercise, we confirm that the embeddings capture content variation: average within-topic cosine similarity is 0.74, compared to 0.15 across topics, and essays responding to the same prompt are more similar to each other (0.78) than essays on the same topic but a different prompt (0.71).

[12] 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.

[13] Pangram's plagiarism checker matches sentences against the open web—webpages, books, news articles, and other publicly indexable sources (Masrour, 2025).

4 Immediate Effects of AI Access on Learning

4.1 First Stage: AI Adoption and Usage Patterns

We begin with the first-stage estimates of how random assignment to AI access translated into AI use. Table 3 reports estimates from equation (1) using several measures of AI usage as outcomes. Panel A presents the revealed-preference measure based on activity in treated students' ChatGPT accounts. Panel B presents 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 $\hat{\beta} = 67.3$ pp as measured by activity in the provided ChatGPT account ($p < 0.001$). This effect represents a large increase from the near-zero baseline in the control group. Self-reported measures yield similar though slightly attenuated effects: any generative AI use increased by $\hat{\beta} = 59.9$ pp ($p < 0.001$). This adoption was almost entirely ChatGPT: a 60.2 pp increase in ChatGPT use ($p < 0.001$), versus just a 1.9 pp increase in the use of other AI models ($p = 0.432$).[14] Consistent with this strong first stage, 88 percent of AI users reported that AI was somewhat or very helpful for the learning task.[15]

Treated students use AI for a mix of purposes—most often to explain concepts, but also to draft and edit text. Figure 3 reports the fraction of treated students who used AI for each of six purposes, measured through self-reports (Panel A) and ChatGPT conversation logs (Panel B), which we classify using LLMs (see Appendix B.6). Both sources agree that explaining concepts is the most common use (46 percent by self-report, 57 percent in the logs), followed by drafting responses and summarizing readings. The ChatGPT usage logs show higher rates of drafting and editing than the self-reports, which suggests that students underreport text-generation uses.[16]

[14] The near-exclusive use of ChatGPT is consistent with surveys of adoption among college students (Hirabayashi et al., 2024; Stöhr et al., 2024; Contractor and Reyes, 2026) and the U.S. working-age population (Bick et al., 2026).

[15] Appendix Table A1 reports correlates of AI take-up among treated students. Frequent prior AI use is the strongest and most robust predictor; self-assessed proficiency and early adoption also predict take-up in bivariate specifications. Take-up is also higher among racial minorities, as white students are significantly less likely to use AI. We find no significant association between take-up and either perceived benefits or GPA; men are directionally more likely to adopt than women, but the difference is imprecise.

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

4.2 Effects on Test Scores

AI access improves students' test performance but not their self-assessed knowledge (Table 4, columns 1–3). Both treated and control students perform better after the learning phase, consistent with skill accumulation. Self-assessed knowledge in the control group rises sharply, from 1.56 to 4.79 on the 0–10 scale ($p < 0.001$). The rise is nearly identical among treated students, leaving an ITT of $\hat{\beta} = 0.02$ points, which is statistically indistinguishable from zero. Test performance, by contrast, responds to AI access. Control students' fraction of questions answered correctly rises from 30.6 percent to 56.3 percent ($p < 0.001$); AI access adds a further $\hat{\beta} = 6.7$ pp, a 12 percent improvement over the control mean. In standardized units, this is a $\hat{\beta} = 0.27$ SD effect (column 2, $p = 0.034$). The corresponding TOT estimate is 10.0 pp, or 0.40 SD ($p = 0.036$, column 3). These effects are similar across the three topics (Appendix Table A2) and robust to excluding students who failed attention or comprehension checks (Appendix Table A3).

The learning gains from AI access are concentrated among middle-performing students. Students with AI access are 8.0 pp more likely to score at least 40 percent correct ($p = 0.098$) and 9.6 pp more likely to score at least 60 percent correct ($p = 0.133$), though both estimates are imprecise (Appendix Table A4). AI access has minimal effects on reaching the 80 percent threshold or achieving perfect scores. Figure 4 illustrates these effects: 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's gains tend to concentrate in the lower half of the performance distribution (Noy and Zhang, 2023; Choi et al., 2024; Doshi and Hauser, 2024; Brynjolfsson et al., 2025; Kanazawa et al., 2025; Dell'Acqua et al., 2026).

Benchmarking the Test-Score Effect. The effect of AI access on Session One test scores is large by the standards of the educational-intervention literature. Kraft (2020) reports a median effect of 0.10 SD among 747 RCTs; our ITT of $\hat{\beta} = 0.27$ SD is more than double this median. Two in-sample benchmarks make this magnitude concrete. A single learning session raises AI-forbidden students' test scores by 1.02 SD above baseline, and AI access adds roughly another quarter of that gain. Among control students, a one-SD increase in college GPA predicts about 0.22 SD higher test performance, so AI access is comparable to a one-SD increase in GPA. Our estimate is similar to the 0.29 SD pooled gain from structured human tutoring programs (Nickow et al., 2024), with the advantage that AI access is not constrained by the supply of qualified tutors. Producing a similar effect

through school spending would cost roughly $8,600 per student over four years (Jackson and Mackevicius, 2024).[17]

Figure 5 places our estimates alongside 22 others from 13 randomized experiments (see Appendix B.7 for the studies and inclusion criteria). The random-effects grand mean is 0.18 SD. The estimates span a wide range, and the spread tracks the role AI plays during practice. The losses often come from settings where AI could do the practice in the learner's place: high-school students with base GPT-4 solve more practice problems yet score −0.19 SD on a later unassisted exam (Bastani et al., 2025). The gains often come from designs that cast AI as a coach: teacher-guided classroom practice with a chatbot raises test scores by 0.21–0.26 SD (De Simone et al., 2025; LearnLM Team, 2026), an AI tutor grounded in the course text raises exam performance by 0.34 SD (Fischer et al., 2025), and guided AI study outperforms unguided access (Hou et al., 2026). Students in our sample receive an unrestricted configuration, yet their gains are positive and persistent. Section 5.3 traces this to how students use AI—whether to augment their effort or automate it.

4.3 Effects on Essay-Based Outcomes

We examine how AI access affects students' writing in three steps. First, we check whether AI traces appear in treated students' essays. Second, we document changes in writing style: length, readability, lexical diversity, and textual similarity. Third, we ask whether AI access improves essay quality. Because Session One essays are written during the learning phase—when treated students have AI access—these estimates reflect three channels: direct AI output, changes in how students approach the writing task, and any learning accumulated during the session itself. We isolate the learning channel in Section 5 by examining Session Two essays, which students write without AI.

AI Detection and Plagiarism. AI use is detectable in treated students' essays (Figure 6 and Table 5, Panel A). The fraction of text classified as AI-generated rises by $\hat{\beta} = 12.3$ pp in the AI-allowed group ($p = 0.019$), a 96 percent increase over the control mean of 12.8 percent.[18] The corresponding TOT estimate is 18.3 pp ($p = 0.017$). We find no evidence of

[17] Jackson and Mackevicius (2024) estimate that increasing per-pupil spending by $1,000 for four years raises test scores by 0.031 SD. Linearly extrapolating, producing an effect of $\hat{\beta} = 0.266$ SD requires $\$1{,}000 \times \hat{\beta}/0.031 \approx \$8{,}600$ per student. While this comparison is necessarily rough, it suggests that AI access can produce learning gains at a fraction of the cost of conventional education spending.

[18] That 12.8 percent of control-group text is flagged as AI-generated likely reflects a combination of control students using AI, false positives, and students whose natural writing style resembles AI-generated

copying from online sources: the fraction of text flagged by Pangram's plagiarism checker is near zero in both groups, with no significant treatment effect.

Linguistic Features. Treated students' essays trend longer and become easier to read, though none of these style effects is statistically significant (Figure 6 and Table 5, Panel B). The length index rises by $\hat{\beta} = 0.13$ SD (column 2). Unpacking the index components, treated students write about 18 more tokens on average, a 9 percent increase over the control mean of 203 tokens (Appendix Table A5, Panel A).[19] The readability index rises by $\hat{\beta} = 0.06$ SD, reflecting similar movements across its components: the Flesch Reading Ease score rises by 4.2 points and mean sentence length falls by 4.7 words (Appendix Table A5, Panel B). The lexical diversity index shows a small positive effect of $\hat{\beta} = 0.05$ SD (column 2). Taken at face value, these estimates suggest AI-driven shifts toward clearer, more accessible writing.

Despite concerns that AI homogenizes writing (Doshi and Hauser, 2024; Meincke et al., 2025; Moon et al., 2025, 2026), we see no such effect in our data (Table 5, Panel C). Treatment effects on cosine similarity are near zero and statistically insignificant: within-group similarity changes by $\hat{\beta} = 0.001$, a 0.1 percent change relative to the control mean of 0.774. Similarity to the provided reading material changes by $\hat{\beta} = -0.004$, a 0.5 percent change relative to the control mean of 0.746. These null results contrast with the convergence patterns documented by Brynjolfsson et al. (2025) among customer-support agents and with the reductions in collective creative diversity observed in lab studies of AI-assisted writing and brainstorming (Doshi and Hauser, 2024; Meincke et al., 2025; Moon et al., 2025). One possible explanation is task structure: customer-service and ideation tasks have well-defined correct answers or converge on common AI-suggested ideas, whereas essay prompts are open-ended and admit multiple valid responses.

Essay Quality and Dimension-Level Ratings. Essay quality is largely unchanged in Session One (Table 6 and Appendix Figure A4). All five dimensions show positive but imprecise effects. Organization and structure ($\hat{\beta} = 0.25$ points on the raw 0–10 scale, or 0.17 SD, $p = 0.237$), accuracy of content ($\hat{\beta} = 0.20$ points, or 0.17 SD), and writing style and clarity ($\hat{\beta} = 0.20$ points, or 0.15 SD) show the largest gains, while the effects

text—a pattern consistent with recent evidence that exposure to LLMs shifts human communication patterns toward AI-like prose (Yakura et al., 2024).

[19] A token is the basic text unit used by language models to process text, corresponding roughly to a short word or sub-word fragment.

on relevance to the prompt ($\hat{\beta} = 0.12$ points, or 0.08 SD) and use of evidence ($\hat{\beta} = 0.08$ points, or 0.05 SD) are smaller. Overall quality shows a small, if imprecise, positive effect: treated essays score $\hat{\beta} = 0.25$ points higher (or 0.17 SD, $p = 0.245$).[20]

4.4 Mechanisms: How AI Transforms the Production Function of Learning

We examine four mechanisms through which AI access affects learning: total time spent learning, how that time is allocated across activities, the learning experience, and academic integrity.

Time Spent Learning. AI access does not change total time spent on the learning phase (Table 7, Panel A). We measure time in two ways: Qualtrics records the duration of the learning phase, and the post-learning survey provides a self-reported measure. Control students spend an average of 32.6 minutes by the Qualtrics measure and 34.3 minutes by self-report. Treatment effects are small and not statistically significant: a 0.2-minute decrease by Qualtrics and a 0.5-minute decrease by self-report. This null contrasts with Noy and Zhang (2023), who find that ChatGPT access reduces task completion time by 0.80 SD among knowledge workers, and with similar time savings documented in other workplace settings (Peng et al., 2023; Dell'Acqua et al., 2026). One possible reason is that workers in those settings often automate task production directly by pasting ChatGPT responses as final output or accepting Copilot's tab completions. Students in our sample, by contrast, primarily use AI for explanation rather than text generation.

Time Allocation Across Learning Activities. Although total learning time is unchanged, its composition shifts away from producing text (Table 7, Panel B). Treated students spend about 5.3 pp less of their learning time on writing activities (drafting, editing, note-taking, and organizing), down from a control-group share of 53.4 percent ($p = 0.028$), and about 4.4 pp more on research activities (reading and searching), up from 44.5 percent ($p = 0.050$). The activity-level breakdown is consistent but individually imprecise: drafting falls the most as a share of time, and note-taking and organizing also decline, while editing and reading rise (Appendix Figure A6). The dominant shift is away from generating text and toward reviewing and absorbing it, with editing the one writing activity that rises, which plausibly reflects students revising AI-generated text rather

[20] Appendix Table A6 reports dimension-level estimates separately for human and AI graders in raw points. Appendix Figure A4 reports the same estimates in SD units.

than drafting their own. This pattern is consistent with a broader finding in professional settings: AI shifts effort from task execution to task stewardship (Lee et al., 2025). For example, Mozannar et al. (2024) find that GitHub Copilot users spend more than half their coding time verifying and editing AI-generated suggestions rather than writing code.

Learning Experience. One potential concern about AI is that it makes learning feel mechanical and alienating, and thus reduces student engagement. We find the opposite. AI access makes learning more enjoyable (Table 7, Panel C): enjoyment of the learning task rises by $\hat{\beta} = 0.66$ points on the 0–10 scale ($p = 0.031$), a 13 percent improvement over the control mean of 5.22; the likelihood of reporting above-median enjoyment rises by 14.1 pp ($p = 0.038$). Effects on how skilled or effective students feel are positive but small: perceived effectiveness on the continuous scale is essentially zero ($\hat{\beta} = 0.13$ points, $p = 0.649$), though the above-median measure rises by 12.1 pp ($p = 0.078$). 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 knowledge workers (though the latter is not statistically significant).

Academic Integrity. AI access increases the likelihood of rule violations on the test (Table 7, Panel D). We examine three measures: proctor-observed cheating, self-reported cheating, and an indicator for any cheating detected by either method. Proctor-observed violations rise by $\hat{\beta} = 5.7$ pp ($p = 0.078$), more than doubling the control mean of 2.9 percent. Self-reported violations rise by $\hat{\beta} = 8.6$ pp ($p = 0.012$), more than quadrupling the control mean of 2.0 percent. On the combined measure, treated students are $\hat{\beta} = 12.6$ pp more likely to cheat ($p = 0.005$), relative to a control mean of 4.9 percent.[21] A back-of-the-envelope calculation suggests that cheating accounts for at most a modest fraction of the test-score effect: multiplying the $\hat{\beta} = 12.6$ pp treatment effect on cheating by the 17.4 pp test-score gap between cheaters and non-cheaters yields roughly 2.2 pp, or about a third of the ITT, under the assumption that the entire score gap between cheaters and non-cheaters reflects cheating itself.

[21] This finding is consistent with student perceptions documented by Ravšelj et al. (2025): 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.

5 Learning Retention and Heterogeneity

AI access raises test scores during Session One. We next examine Session Two outcomes measured about one week later without AI to assess whether the gains from AI-assisted learning persist when AI is removed.[22]

5.1 Persistence of Effects on Test Scores

The learning gains from AI access persist one week later (Table 4, columns 4–6). Both treated and control students perform worse in Session Two than in Session One, consistent with skill depreciation. Self-assessed knowledge in the control group falls from 4.79 to 3.83 on the 0–10 scale ($p < 0.001$). The decline is smaller among treated students, with an ITT of $\hat{\beta} = 0.18$ points, though this difference is small and not statistically distinguishable from zero. Test performance shows a similar fade-out pattern. Control students' fraction of questions answered correctly falls from 56.3 percent in Session One to 49.5 percent in Session Two ($p = 0.008$). Treated students decline as well but remain ahead of the control group. The resulting ITT of $\hat{\beta} = 5.1$ pp ($p = 0.027$) is about 76 percent of the 6.7 pp Session One effect (though, given the standard errors, the two effects are not statistically distinguishable). Standardized, this corresponds to a $\hat{\beta} = 0.27$ SD effect.[23]

Session Two learning gains appear in the middle of the score distribution (Appendix Table A4, columns 4–6). Treated students are $\hat{\beta} = 12.2$ pp more likely to score at least 60 percent correct ($p = 0.051$). Effects at the other thresholds are smaller and statistically insignificant. Panel B of Figure 4 illustrates these patterns. Across sessions, the test-score gains remain in the middle of the distribution, echoing the Session One pattern. AI-assisted learning therefore produces durable, though partially decaying, knowledge gains.

5.2 Persistence of Effects on Essay-Based Outcomes

AI Detection and Plagiarism. AI traces are minimal in Session Two, consistent with students writing without AI access (Table 5, Panel A). The treatment effect collapses from

[22]Students were not informed that they would complete additional assessments in Session Two. We find no evidence that students prepared between sessions: only 5 percent of students reported looking up extra information about the topic, and fewer than 1 percent reported studying it, with no significant differences between treatment and control groups ($p = 0.196$ and $p = 0.319$, respectively).

[23]This retention effect is consistent with the two other studies that measure learning retention, both of which report positive effects: Lira et al. (2025) estimates 0.41 SD on a 24-hour retention test of cover-letter writing, and Kazemitabaar et al. (2023) estimates 0.41 SD on a one-week code-modification test.

$\hat{\beta} = 12.3$ pp to $\hat{\beta} = 0.0$ pp ($p = 0.991$). Plagiarism remains near zero in both groups, with no significant treatment effect.

Linguistic Features. Effects on essay style fade in Session Two (Table 5): most Session Two coefficients are smaller than the Session One estimates. The length index falls from $\hat{\beta} = 0.13$ SD in Session One to $\hat{\beta} = 0.05$ SD in Session Two. Readability remains at $\hat{\beta} = 0.06$ SD. Lexical diversity reverses direction, falling from 0.05 SD to -0.13 SD. The widespread fade-out suggests that the modest Session One style differences arise from AI writing entering students' essays rather than from durable changes in how students write: the differences appear when students have AI and vanish when they do not.

Essay Quality. AI access improves essays in Session Two, when students write without AI (Figure 6 and Table 6, columns 4–6). Writing style and clarity and relevance to the prompt both show statistically significant gains ($\hat{\beta} = 0.41$ points, or 0.30 SD, $p = 0.016$; and $\hat{\beta} = 0.41$ points, or 0.26 SD, $p = 0.041$, respectively). The remaining dimensions—accuracy, evidence, and organization and structure—show positive but imprecise effects (with p-values of 0.101, 0.111, and 0.240, respectively). Overall quality rises by $\hat{\beta} = 0.31$ points (or 0.20 SD), though this holistic rating is imprecisely estimated ($p = 0.143$). These effects suggest that AI access improves students' higher-order learning, not just how many facts they recall.

5.3 Automation versus Augmentation

AI is a general-purpose technology with many uses, so the learning effects of AI—and whether they persist over time—may depend on how students use it. One 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). Automation may reduce the cognitive effort students invest in learning. Augmentation may instead raise the productivity of that effort.

To identify augmentation versus automation empirically, we asked an LLM to read each ChatGPT conversation log and label it *Automation* if the AI does the work *for* the student, *Augmentation* if the AI works *with* the student, *Mixed* if AI does both, or *Other* if the conversation is off-topic (see Appendix B.6 for the full prompt). Of treated AI users, 49 percent are pure augmentation users, 32 percent are pure automation users, 8 percent are mixed, and the remaining 11 percent are classified as "other," having engaged AI only

off-topic.[24]

Three pieces of evidence validate this classification. First, the two groups prompt AI differently: automation users rely on AI more for drafting and editing, while augmentation users turn to AI more for explaining concepts (Appendix Figure A7). Second, automation users reallocate time during the learning phase: relative to augmentation users, they spend about 17 percent less time on research activities (reading and searching; $p = 0.034$) and 8 percent less time overall ($p = 0.045$). Third, these behavioral differences show up in essay outputs: Pangram flags 53.6 percent of automation users' text as AI-generated, versus 20.9 percent for augmentation users.

Automation and augmentation users show sharply different patterns of effects (Table 8). In Session One, the effect of AI access on essay quality is large for automation users but small for augmentation users ($\hat{\beta} = 0.54$ SD, $p = 0.029$, versus $\hat{\beta} = 0.05$ SD on overall quality), while the effect on test scores is slightly smaller for automation than for augmentation users ($\hat{\beta} = 0.33$ SD versus $\hat{\beta} = 0.44$ SD)—consistent with AI doing the writing rather than teaching the student. By Session Two—when students write without AI access—the patterns diverge. The effects on essay quality for automation users fade out entirely, with point estimates indistinguishable from zero ($\hat{\beta} = 0.02$ SD on overall quality, $\hat{\beta} = 0.04$ SD on the quality index), and their test-score effect attenuates to $\hat{\beta} = 0.19$ SD ($p = 0.330$). Augmentation users, in contrast, retain a positive—though imprecise—essay-quality effect in Session Two ($\hat{\beta} = 0.22$ SD on overall quality, $p = 0.216$) and show large test-score gains ($\hat{\beta} = 0.29$ SD, $p = 0.061$). In short, automation's Session One advantage reflects AI-produced output and does not survive its removal, whereas augmentation's gains persist unaided, consistent with skill accumulation.

These findings are consistent with Strömberg et al. (2026), who find that AI raises homework scores but lowers exam performance, with the learning losses concentrated among students whose unusually short completion times and high homework scores indicate they outsource their work to AI—the field analogue of our automation users. This divergence matches a recurring pattern in the AI-and-learning literature: AI access can boost short-term output while reducing what students learn from the task (Bastani et al., 2025; Liu et al., 2026; Shen and Tamkin, 2026; Strömberg et al., 2026). The contrast between our augmentation and automation users shows that this tradeoff is not inevitable: when stu-

[24]Similar mixed usage patterns appear in observational data: Contractor and Reyes (2026) in survey responses from Middlebury undergraduates, Handa et al. (2025) and OpenAI (2025) in large-scale Claude and ChatGPT logs, and Ammari et al. (2025) in ChatGPT logs from undergraduates at another U.S. university.

dents use AI to scaffold their cognitive effort rather than substitute for it, learning gains can persist.

5.4 Heterogeneity of Treatment Effects

Does AI access widen or narrow pre-existing gaps among students? To assess this, we examine heterogeneity along several dimensions: academic ability (GPA and SAT quartiles), prior AI experience, and demographic characteristics (gender, race, nationality, and field of study).

Academic Ability. AI access raises test scores more among higher-ability than lowest-ability students (Appendix Figure A8). Gains are smallest in the bottom quartile of either measure ($\hat{\beta} = 0.06$ SD for SAT and $\hat{\beta} = 0.05$ SD for GPA) and larger in the upper quartiles: the SAT gain peaks in the third quartile ($\hat{\beta} = 0.40$ SD), and the GPA gain is spread across the upper three quartiles (0.19–0.29 SD). This suggests that AI access may widen gaps relative to the lowest-ability students.

AI Experience and Demographics. Neither prior AI experience nor demographic characteristics systematically moderate the test-score effect (Appendix Figure A9 and Appendix Table A7). We find no evidence that essay-quality gains systematically vary with students' self-assessed AI proficiency or frequency of use. We find suggestive differences by gender (women gain 0.46 SD more than men on Session One tests, $p = 0.079$) but no systematic differences by race, nationality, or field of study.

6 Beliefs About AI's Impact on Learning

Students' beliefs about AI's effect on learning may shape their adoption decisions. In observational data, students report that AI improves their learning (Stöhr et al., 2024; Ravšelj et al., 2025). Yet these beliefs may not track actual effects: students hold biased beliefs in other educational settings (Jensen, 2010; Wiswall and Zafar, 2015), and even experienced professionals misjudge AI's productivity effects.[25]

We test the accuracy of students' beliefs—and how treatment shapes them—in two ways. First, after the Session Two test, students estimate (1) how many questions they

[25] Becker et al. (2025) find that experienced open-source developers predicted AI tools would speed them up by 24 percent and, after the study, still believed AI had accelerated their work by roughly 20 percent. The actual measured effect was a 19 percent *slowdown*.

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. From these we compute perceived treatment effects on own and others' performance and compare them with the actual estimates from Section 5, overall and across subgroups. Second, we analyze responses to an open-ended question about the effects of generative AI on student learning in college. We validate students' open-ended responses against their AI usage patterns (Appendix C.1) and, following Andre et al. (2026), code each response as a causal graph of the mechanisms it describes (Appendix C.2).

6.1 Beliefs About AI's Effect on Test Scores

Both groups believe AI improves test performance, but only treated students gauge the magnitude correctly (Figure 7, Panel A, and Appendix Table A8). We examine two measures: the perceived AI gain in the student's own performance and in others', both measured in percentage points. The actual ITT, for reference, is $\hat{\beta} = 5.1$ pp. Both groups give similar predictions of their own raw score (control 37.5 percent correct, treated 37.2 percent, both below the actual mean of 51.0 percent); what differs is what they attribute to AI. Control students predict an AI gain for themselves of 25.2 pp, about five times the actual effect; treated students, who experienced AI firsthand, predict only 3.8 pp, close to the effect we estimate.[26] Predictions about others' performance show the same pattern in muted form: control students predict 14.1 pp and treated 10.6 pp, both still inflated relative to the actual effect. Beliefs also track actual gains across subgroups: the subgroups that benefit most from AI perceive the largest gains (Figure 7, Panel B).

6.2 Students' Mental Models of AI and Learning

The above beliefs capture students' expected test-score gains from AI, not the mechanisms by which they think AI affects learning. To recover these mental models, we asked an open-ended question on the exit survey: "In your opinion, how does generative AI (e.g., ChatGPT) affect student learning in college? Please explain your reasoning." Following Andre et al. (2026), we code each short narrative as a causal graph running from AI use to learning through the mechanisms the student names (see Appendix C.2 for the

[26] The lower average perceived effect among treated students reflects two shifts in the distribution of beliefs relative to control students (Appendix Figure A10): more mass at zero (no perceived effect) and less mass at large positive values (five or more questions).

coding procedure and its reliability, and Appendix C.3 for additional results). Figure 8 aggregates students' mental models into a single causal graph in which each node is one such mechanism (sized by how often students name it), and each link is green where they describe the mechanism as promoting learning and maroon where it harms it.

The augmentation-versus-automation distinction that moderates our treatment effects is also the structure most students use to reason about the effects of AI on learning. Among students naming any learning channel, 69 percent name both one through which AI helps and one through which it harms. The single most common element, present in 54 percent of coded responses, is that the effect depends on how AI is used; as one student put it, AI is "a double-edged sword, with each individual student's use determining whether it is helpful or hurtful for their education." The two most common mechanisms map onto the two sides of this distinction: AI explaining concepts and tutoring, as an augmentation channel (40 percent of coded responses), and AI shortcutting the work, as an automation channel (42 percent).[27] Taken together, students' reasoning mirrors our estimates: they expect AI to raise their learning—accurately so once they have used it—and they locate the gains where our treatment effects concentrate, in AI that augments their work rather than does it for them.

7 Conclusion

This paper provides causal evidence on how access to generative AI affects student learning. In a randomized experiment with undergraduate students, AI access produces learning gains that, on average, persist one week later. Gains are concentrated in the middle of the score distribution on both the immediate and retention tests. AI access also reshapes how students approach learning a new topic: they spend less time drafting text and more time reading and searching for information, and report higher enjoyment. Taken together, these results provide proof of concept that off-the-shelf AI can improve learning, with effects that depend on how students use it.

We do not view our findings as implying that widespread AI adoption will raise learning overall. Our estimates capture learning per unit of time. We find no effect on total learning time—likely because the lab setting offered few competing uses of time—so the learning

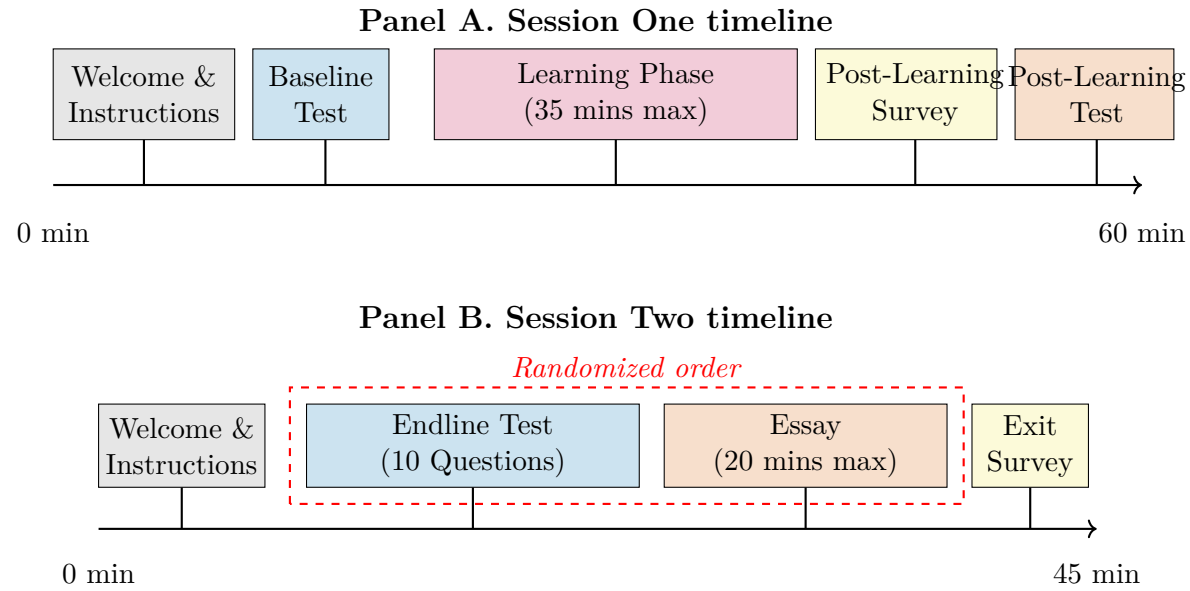
[27] Appendix Table C1 reports how students' narratives about AI and learning vary with their characteristics. Self-assessed AI proficiency is the strongest predictor of framing: conditional on demographics and frequency of use, above-median-proficiency students are more likely to frame AI as augmentation and less likely to frame it as automation.

gains hold time-on-task fixed. In ordinary academic work, students choose how long to spend on each task, and many use AI to save time. Whether AI raises or lowers total learning therefore depends on how students reallocate the time they save and on whether productivity gains per unit of time more than compensate for any reduction in time spent learning.

Our findings have implications for incentive design in higher education. Augmentation uses are more likely to raise learning than automation uses, but how students choose between them is endogenous to the incentives they face. For example, grade inflation weakens the link between effort and grades, pushes students to differentiate themselves through extracurriculars or other non-academic activities, and gives them reason to automate coursework with AI to free up time. Similarly, students who view college primarily as a signaling mechanism rather than as human-capital accumulation have weaker incentives to use AI as a learning tool, since learning itself plays a smaller role in their perceived returns to education. How higher-education incentives shape AI usage—and, through it, learning—is a promising avenue for future research.

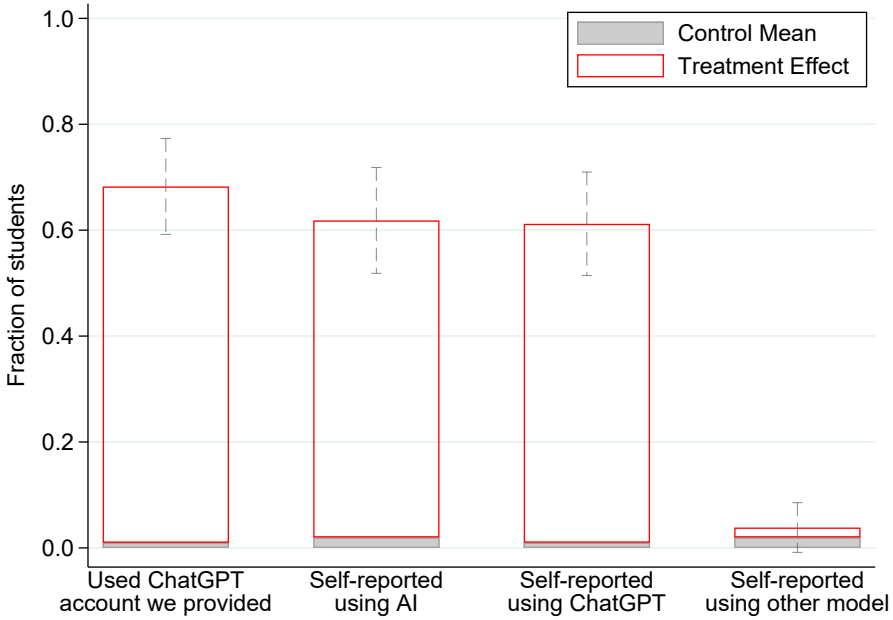
Figures and Tables

Figure 1: Experimental Sessions Timelines



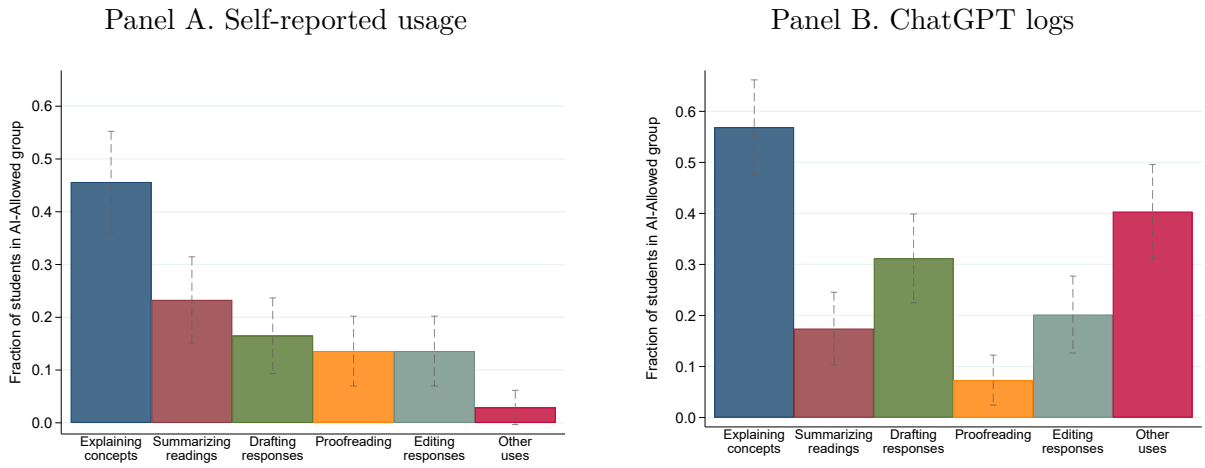
Notes: This figure shows the timeline of the two experimental sessions. Panel A shows the structure of Session One, during which students were randomly assigned to AI-allowed or AI-forbidden conditions. Panel B shows Session Two, conducted approximately one week later (mean of 6.99 days), during which all students 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 students. The Baseline Test and Post-Learning Test in Session One each contained 5 multiple-choice questions. The Endline Test in Session Two contained 10 multiple-choice questions.

Figure 2: The Impact of AI Access 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 is constructed from activity in the ChatGPT account provided to each student. The remaining three measures are based on self-reported usage collected in the exit survey. Vertical bars represent 95 percent confidence intervals.

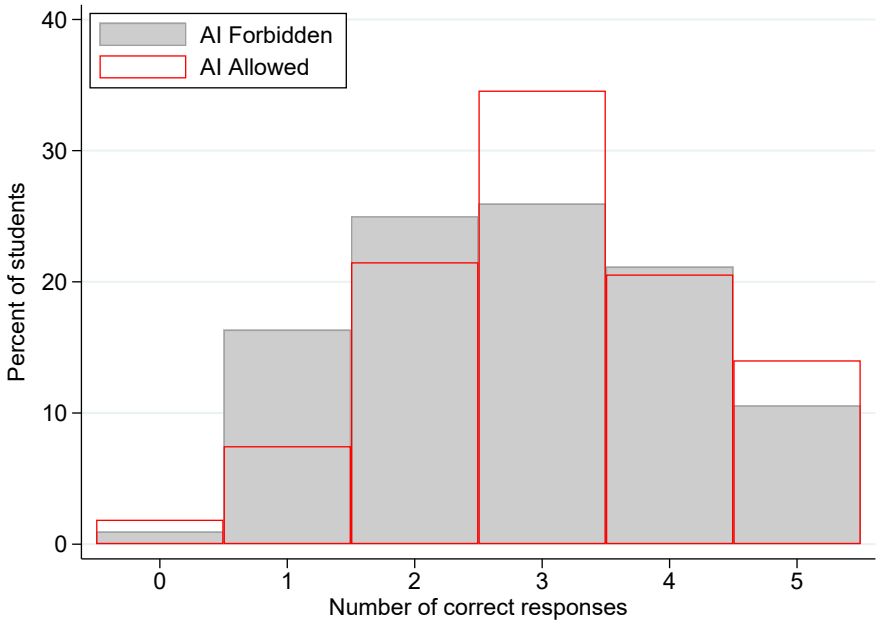
Figure 3: Types of Generative AI Use During the Learning Phase



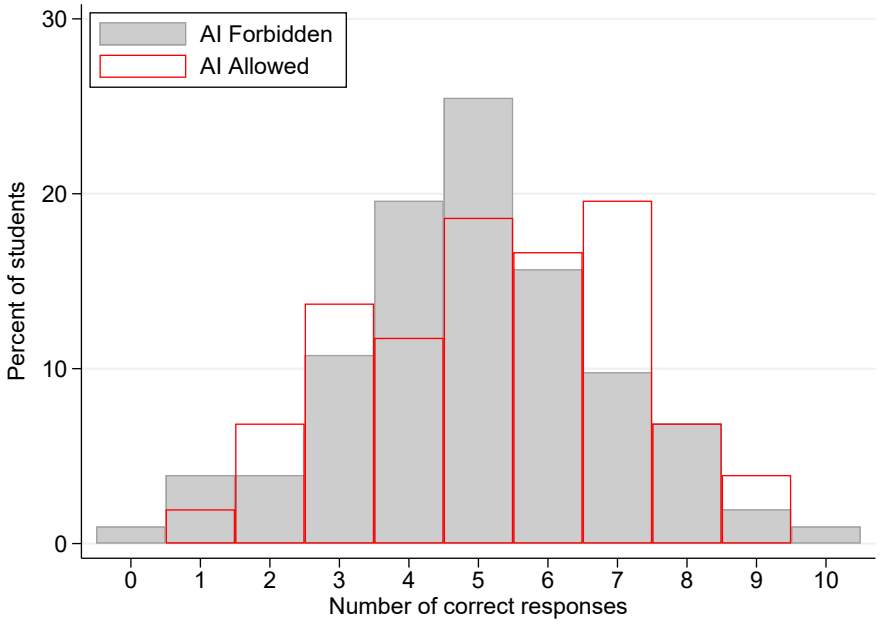
Notes: This figure shows AI usage types among treated students. Panel A presents self-reported usage types collected in the exit survey; students could select multiple categories. Panel B is constructed from the actual ChatGPT conversation logs. We first use LLMs to classify each student prompt into one of six categories: explaining concepts, summarizing readings, drafting responses, proofreading, editing responses, and other uses (see Appendix B.6 for the classification procedure). We then construct student-level indicators: for each category, the indicator equals one if any of the student's prompts fell into that 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.

Figure 4: Distribution of Test Performance by Treatment Group

Panel A. Session One (knowledge acquisition)

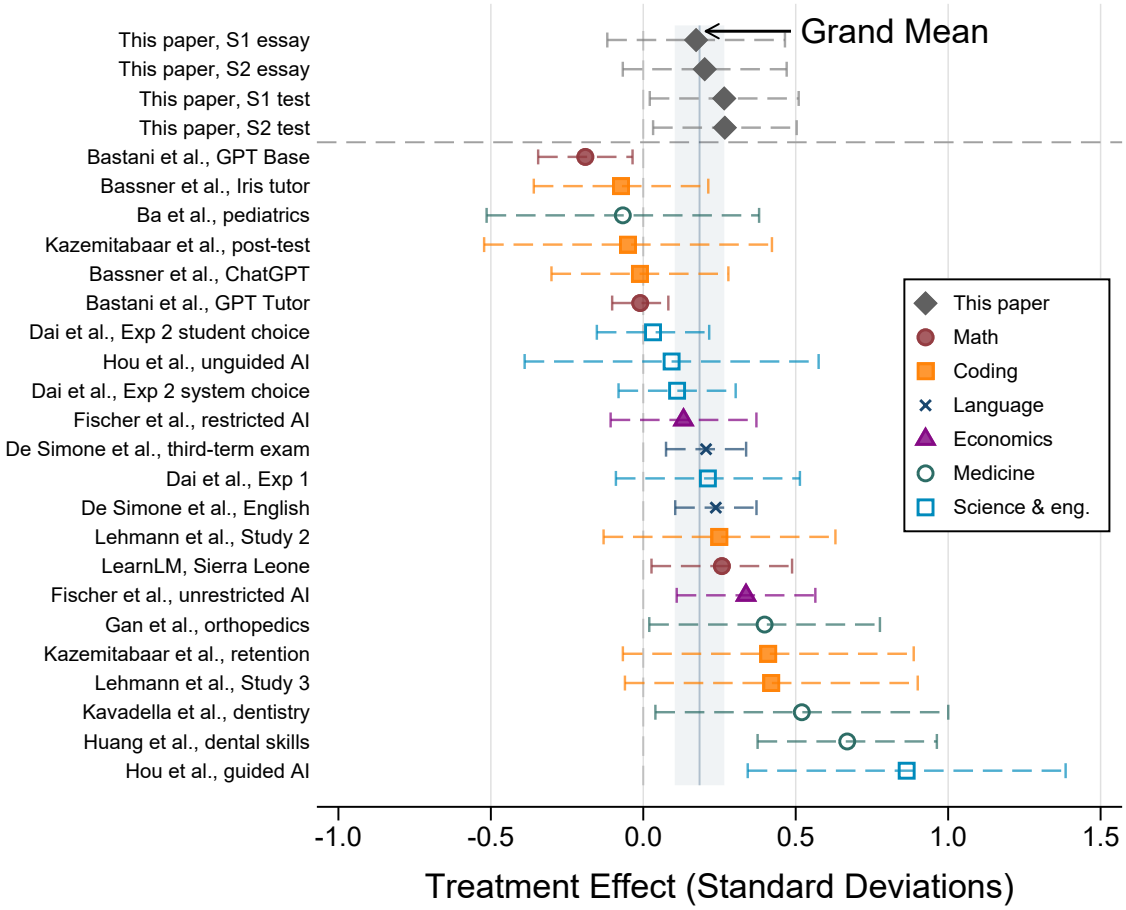


Panel B. Session Two (knowledge retention)



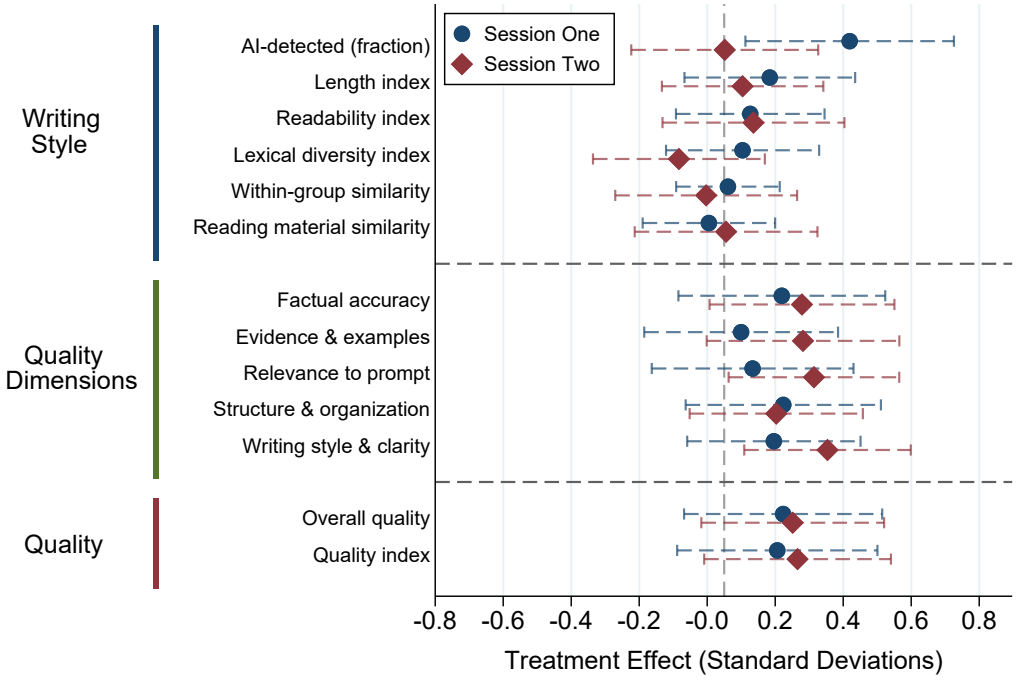
Notes: This figure shows the distribution of test scores by treatment assignment. The x-axis shows the number of correct responses; the y-axis shows the percentage of students achieving each score. The Session One test had 5 questions; the Session Two test had 10 questions.

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



Notes: This figure compares treatment effect sizes (in standard deviations) across randomized 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 this paper, circles for math, squares for coding, crosses for language learning, triangles for economics, hollow circles for medicine, and hollow squares for science and engineering. 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 come from the 13 randomized experiments described in Appendix B.7 and summarized in Appendix Table B3. Horizontal lines show 95 percent confidence intervals. The shaded band and vertical line show the random-effects grand mean and its 95 percent confidence interval, estimated via the DerSimonian and Laird (1986) method. The dashed vertical line marks zero.

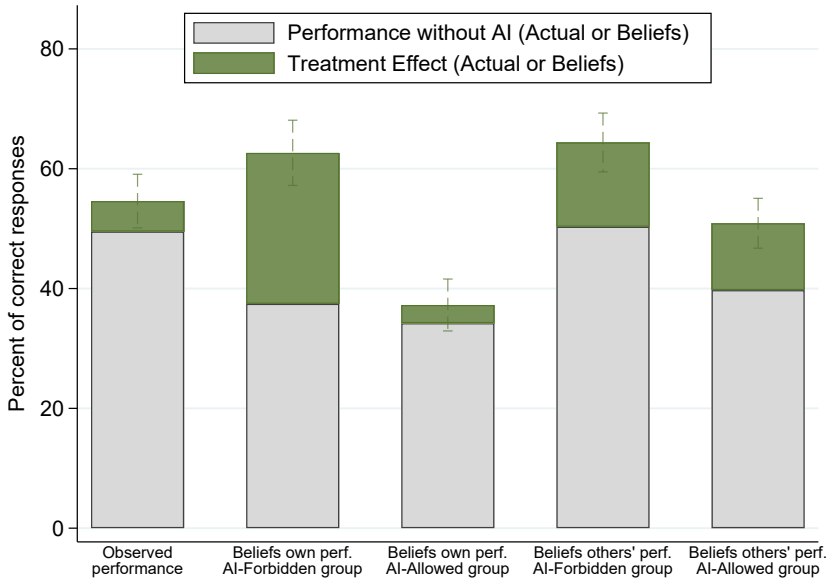
Figure 6: Effects of AI Access on Essay Quality and Linguistic Features



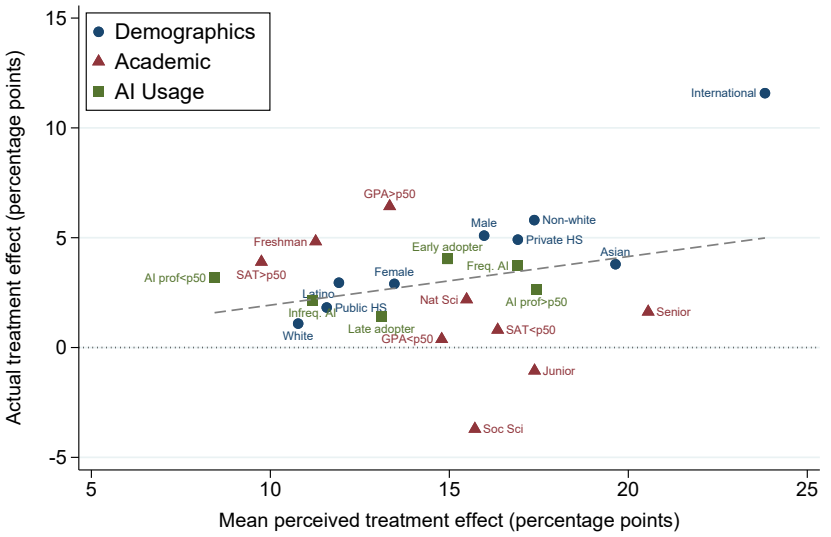
Notes: This figure presents treatment effects of AI access on writing style (top panel), individual quality dimensions (middle panel), and overall essay quality (bottom panel), measured in standard deviations. The writing style indices are: the AI-detected fraction (the share of text classified as AI-generated by Pangram, standardized in the control group); 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); within-group similarity (the average pairwise cosine similarity within treatment×topic×prompt cells); and reading material similarity (between the student essay and the provided reading material). The quality dimensions are the five sub-components of a standardized rubric—accuracy of content, use of evidence and examples, relevance to the prompt, organization and structure, and writing style and clarity—each averaged across the human and AI graders. The overall quality measures are overall quality (the grader's holistic assessment) and the quality index (averaging the five dimensions above), each averaged across the human and AI graders. 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. Point estimates underlying each coefficient in raw units are reported in Tables 5 and 6. The five quality dimensions split by grader appear in Appendix Figure A4, and the individual essay characteristics underlying the writing style indices appear in Appendix Figure A5.

Figure 7: Actual and Perceived Treatment Effects on Test Performance

Panel A. Beliefs versus actual effects

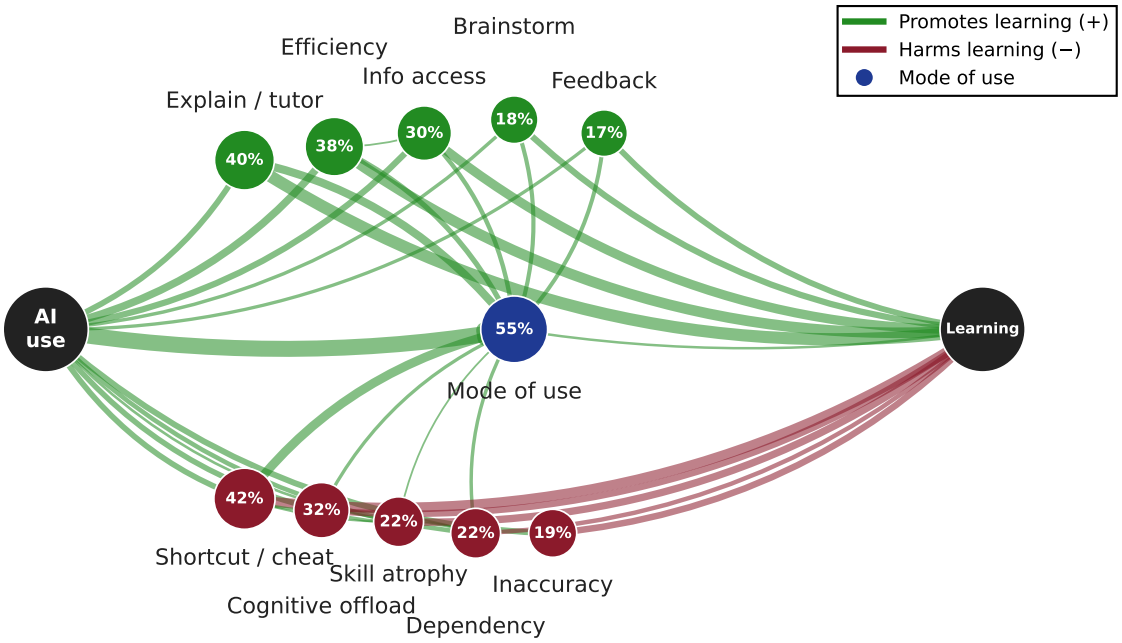


Panel B. Perceived versus actual effects by subgroup



Notes: This figure shows actual and perceived treatment effects of AI access on test performance. In Panel A, the first bar shows observed performance—gray is the control mean, and green is the estimated treatment effect. The remaining bars show students' beliefs about own and others' performance, separately by treatment group. Panel B plots the mean perceived treatment effect (own performance) against the actual treatment effect for each demographic, academic, and AI-usage subgroup with at least 40 observations; the dashed line is a linear fit. Vertical bars represent 95 percent confidence intervals.

Figure 8: The Average Narrative About AI's Effect on Learning



Notes: This figure shows the aggregate signed causal graph across the 200 open-ended responses codable as a narrative, constructed following Andre et al. (2026). We code each student's answer to the question "in your opinion, how does generative AI (e.g., ChatGPT) affect student learning in college? Please explain your reasoning" as a causal graph from AI use (left) to learning (right) through the mechanisms the student names; the figure overlays these graphs. Node size is proportional to the share of students who name a mechanism, and link width to the share drawing the link; we omit mechanisms below 10 percent and links below 2 percent. Green links denote mechanisms described as promoting learning, maroon links mechanisms that harm it, and navy the mode-of-use node. Appendix C.2 describes the coding and its reliability.

Table 1: Summary Statistics

	Signed-up (1)	Attended Session One (2)	Attended Both Sessions (3)
Panel A. Demographic characteristics			
Age	20.211	20.147	20.132
Male	0.359	0.351	0.353
White	0.496	0.498	0.510
International student	0.270	0.265	0.275
Public high school	0.539	0.540	0.534
Panel B. Academic background			
Took SAT/ACT	0.742	0.763	0.770
SAT score (conditional)	1387.593	1385.969	1386.571
SAT score (including predicted)	1386.802	1385.403	1386.716
College GPA	3.680	3.683	3.686
Freshman	0.359	0.374	0.382
Already declared major	0.780	0.780	0.777
Natural-science major	0.301	0.308	0.314
Social-science major	0.359	0.355	0.348
Humanities/Arts/Languages major	0.109	0.104	0.103
Study hours per week	16.929	16.668	16.686
Panel C. Attendance and comprehension			
Attended Session One	0.824	1.000	1.000
Attended Session Two	0.797	0.967	1.000
Attended Session Two (conditional S1)	0.967	0.967	1.000
Attended both sessions	0.797	0.967	1.000
Days between sessions	6.951	6.986	6.971
Comprehension score (0-5)	4.768	4.768	4.765
Panel D. Experimental conditions			
Assigned Blockchain topic	0.365	0.365	0.368
Assigned CRISPR topic	0.327	0.327	0.324
Assigned Carbon-capture topic	0.308	0.308	0.309
Baseline self-assessed knowledge (0-10)	1.517	1.517	1.520
Fraction correct (baseline)	0.303	0.303	0.310
Shared info between sessions	0.063	0.063	0.064
Looked up info about topic	0.049	0.049	0.049
Studied topic for Session Two	0.005	0.005	0.005
Number of students	256	211	204

Notes: This table presents summary statistics for three groups of students: students who completed the sign-up survey (column 1), those who attended Session One (column 2), and those who attended both sessions (column 3). 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)
	(1)	(2)	(3)	(4)
Panel A. Demographic characteristics				
Age	20.016	20.400	0.382	(0.227)*
Male	0.405	0.315	-0.091	(0.060)
White	0.524	0.469	-0.055	(0.063)
International student	0.286	0.254	-0.033	(0.055)
Public high school	0.540	0.538	-0.001	(0.063)
Panel B. Academic background				
Took SAT/ACT	0.778	0.708	-0.067	(0.054)
SAT score (conditional)	1392.371	1382.554	-7.862	(21.039)
SAT score (including predicted)	1391.418	1382.382	-8.849	(17.999)
College GPA	3.694	3.666	-0.025	(0.038)
Freshman	0.357	0.362	0.007	(0.061)
Already declared major	0.817	0.742	-0.075	(0.052)
Natural-science major	0.302	0.300	-0.002	(0.058)
Social-science major	0.397	0.323	-0.077	(0.061)
Humanities/Arts/Languages major	0.111	0.108	-0.000	(0.039)
Study hours per week	16.864	16.992	0.230	(1.197)
Panel C. Attendance and comprehension				
Attended Session One	0.825	0.823	-0.006	(0.048)
Attended Session Two	0.810	0.785	-0.027	(0.050)
Attended Session Two (conditional S1)	0.981	0.953	-0.026	(0.024)
Attended both sessions	0.810	0.785	-0.027	(0.050)
Days between sessions	6.975	6.929	-0.030	(0.156)
Comprehension score (0-5)	4.769	4.766	-0.009	(0.065)
Panel D. Experimental conditions				
Assigned Blockchain topic	0.375	0.355	-0.023	(0.066)
Assigned CRISPR topic	0.327	0.327	0.002	(0.064)
Assigned Carbon-capture topic	0.298	0.318	0.021	(0.062)
Baseline self-assessed knowledge (0-10)	1.558	1.477	-0.086	(0.251)
Fraction correct (baseline)	0.306	0.301	-0.007	(0.038)
N (Students)	126	130	256	

Notes: This table shows average student characteristics by treatment assignment. Students in the AI-allowed group could use generative AI tools during the learning phase; students in the AI-forbidden group could not. Panels A and B report characteristics measured before treatment assignment. Panel C reports session attendance and instruction comprehension, realized after assignment; Panel D reports the randomly assigned essay topics and baseline knowledge measures collected at the start of Session One, before the learning phase. 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 ChatGPT account	0.010	0.673***
		(0.046)
Number of prompts	0.000	3.909***
		(0.426)
Number of conversations	0.000	0.745***
		(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 tool	0.020	0.019
		(0.024)
N	104	211

Notes: This table reports first-stage effects of AI-access assignment on measures of AI usage. Each row reports the effect of being assigned to the AI-allowed group on the indicated measure. Panel A presents revealed measures: Used ChatGPT account equals one if at least one prompt was recorded in the ChatGPT account provided to the student; 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 Test Scores

	Session One			Session Two		
Outcome	Control mean (1)	ITT (2)	TOT (3)	Control mean (4)	ITT (5)	TOT (6)
Self−assessed	4.788	0.021	0.030	3.833	0.185	0.270
		(0.195)	(0.287)		(0.201)	(0.292)
Fraction correct	0.563	0.067**	0.100**	0.495	0.051**	0.075**
		(0.032)	(0.048)		(0.023)	(0.034)
Test score (SD)	0.000	0.266**	0.396**	0.000	0.268**	0.393**
		(0.125)	(0.188)		(0.120)	(0.181)
N	104	211	211	102	204	204

Notes: This table reports treatment effects of AI access on learning outcomes in Session One (with AI) and Session Two (without AI). 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, with 0 indicating "I know nothing about this topic" and 10 indicating "I am an expert." Fraction correct is measured on a 0–1 scale. Test score (SD) is the fraction correct standardized to have mean zero and standard deviation one 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: Effects of Generative AI Access on Essay Linguistic Features

Outcome	Session One (with AI)			Session Two (without AI)		
	Control mean (1)	ITT (2)	TOT (3)	Control mean (4)	ITT (5)	TOT (6)
Panel A. AI detection and plagiarism						
AI−detected (fraction)	0.128	0.123**	0.183**	0.040	0.000	0.000
		(0.052)	(0.076)		(0.028)	(0.041)
Plagiarism (fraction)	0.000	0.000	0.000	0.000	0.000	0.000
		(0.000)	(0.000)		(0.000)	(0.001)
Panel B. Writing style						
Length index	0.000	0.131	0.199	0.000	0.053	0.078
		(0.125)	(0.190)		(0.118)	(0.174)
Readability index	0.000	0.061	0.090	0.000	0.065	0.097
		(0.088)	(0.132)		(0.103)	(0.153)
Lex. diversity index	0.000	0.054	0.083	0.000	−0.132	−0.195
		(0.114)	(0.177)		(0.128)	(0.190)
Panel C. Homogeneity and similarity						
Within−group sim.	0.774	0.001	0.001	0.762	−0.005	−0.007
		(0.007)	(0.010)		(0.013)	(0.019)
Reading material sim.	0.746	−0.004	−0.006	0.704	0.001	0.001
		(0.009)	(0.014)		(0.013)	(0.019)
N	104	210	210	99	199	199

Notes: This table reports treatment effects of AI access on essay characteristics, including AI-detection and plagiarism measures, writing-style indices, and textual similarity. Each row reports the effect of being assigned to the AI-allowed group on the indicated outcome. Panel A reports AI detection and plagiarism measures: AI-detected (fraction) is the fraction of text classified as AI-generated by the Pangram AI content detector; Plagiarism (fraction) is the fraction of text flagged as plagiarized by Pangram's plagiarism checker. Panel B reports z-scored writing style indices, standardized relative to the control group (Appendix Table A5 reports results for individual writing characteristics). 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 C reports homogeneity and 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; reading material similarity is the cosine similarity between a student's essay embedding and the embedding of the provided reading material. Both are computed using a sentence-embedding model (Reimers and Gurevych, 2019) fine-tuned from the MPNet architecture (Song et al., 2020). 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	ITT	TOT	Control mean	ITT	TOT
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Individual dimensions (human and AI averaged)						
Accuracy	6.745	0.198	0.296	5.204	0.312	0.466
		(0.182)	(0.271)		(0.189)	(0.288)
Evidence	6.143	0.076	0.114	4.095	0.314	0.473
		(0.223)	(0.333)		(0.196)	(0.302)
Relevance	6.957	0.118	0.178	6.073	0.409**	0.597**
		(0.214)	(0.322)		(0.198)	(0.295)
Organization	5.930	0.249	0.382	5.310	0.244	0.355
		(0.210)	(0.322)		(0.207)	(0.305)
Writing style	6.475	0.202	0.304	5.806	0.411**	0.618**
		(0.180)	(0.271)		(0.169)	(0.264)
Panel B. Overall quality						
Overall quality (average)	6.190	0.246	0.373	5.076	0.308	0.459
		(0.211)	(0.321)		(0.210)	(0.317)
Quality index (average)	6.450	0.190	0.289	5.297	0.281	0.419
		(0.183)	(0.279)		(0.183)	(0.276)
N	104	209	209	97	197	197

Notes: This table reports treatment effects of AI access on essay quality, scored across five sub-component dimensions and overall, averaging human and AI grades. 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 individual dimension scores averaged across human and AI graders. Panel B reports overall quality and the average of the five sub-component scores (accuracy, evidence, relevance, organization, and writing style), each averaged across human and AI graders. Both human- and AI-grading regressions are at the student level; human grades are averaged across each essay's graders before estimation. 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). Intent-to-treat (ITT) estimates report the effect of being assigned to the AI-allowed group; treatment-on-the-treated (TOT) estimates instrument ChatGPT use with random 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 7: Mechanisms: Time Use, Experience, and Academic Integrity

Outcome	Control mean (1)	ITT (2)	TOT (3)
Panel A. Time spent learning			
Time spent learning (Qualtrics, minutes)	32.611	−0.205	−0.304
		(0.678)	(1.008)
Time spent learning (self−reported, minutes)	34.312	−0.532	−0.802
		(0.669)	(1.009)
Panel B. Time allocation			
Share on research activities (percent)	44.462	4.435**	6.683**
		(2.245)	(3.362)
Share on writing activities (percent)	53.381	−5.347**	−8.056**
		(2.407)	(3.571)
Panel C. Learning experience			
Enjoyed learning (continuous)	5.221	0.665**	1.008**
		(0.306)	(0.474)
Enjoyed learning (above median)	0.481	0.141**	0.211**
		(0.067)	(0.103)
Found effective (continuous)	5.000	0.129	0.191
		(0.282)	(0.420)
Found effective (above median)	0.529	0.121*	0.181*
		(0.068)	(0.103)
Panel D. Academic integrity			
Caught by proctor	0.029	0.057*	0.085*
		(0.032)	(0.047)
Self−reported AI use	0.020	0.086**	0.126**
		(0.034)	(0.050)
Any integrity violation	0.049	0.126***	0.185***
		(0.044)	(0.064)
N	102	205	205

Notes: This table reports treatment effects of AI access on four mechanisms: total time spent learning (Panel A), how that time is allocated across activities (Panel B), the learning experience (Panel C), and academic integrity (Panel D). Each row reports the effect of being assigned to the AI-allowed group on the indicated outcome. Time on task (Qualtrics) and Time on task (self-reported) are the duration of the learning phase in minutes, recorded by Qualtrics and in the post-learning survey, respectively. The Share on research, writing, and other activities rows express each category's self-reported minutes as a percentage of the student's total self-reported learning time; these rows report effects in percentage points (the control-mean column shows the control-group share). Enjoyed learning and Found effective are 0–10 self-reports; the (above median) rows are the corresponding binary indicators. Caught by proctor, Self-reported AI use, and Any integrity violation are indicators for being flagged by a proctor during the knowledge tests, admitting unauthorized AI use on the exit survey, and either occurring; the knowledge tests prohibited all students from using external resources. All specifications include controls selected through a double-lasso procedure (Belloni et al., 2014). Heteroskedasticity-robust standard errors in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Table 8: Effects of AI Access by Type of AI Use: Automation versus Augmentation

	Session One			Session Two		
	Test score (1)	Overall quality (2)	Quality index (3)	Test score (4)	Overall quality (5)	Quality index (6)
Panel A. All students						
Overall (TOT)	0.396**	0.263	0.238	0.393**	0.300	0.310
	(0.188)	(0.226)	(0.229)	(0.181)	(0.207)	(0.213)
N	211	209	209	204	197	197
Panel B. By type of AI use						
Automation users	0.326	0.536**	0.524**	0.189	0.017	0.036
	(0.229)	(0.242)	(0.241)	(0.193)	(0.237)	(0.246)
N	133	132	132	131	126	126
Augmentation users	0.444***	0.055	0.037	0.292*	0.220	0.249
	(0.169)	(0.191)	(0.193)	(0.154)	(0.177)	(0.186)
N	147	147	147	143	136	136

Notes: This table reports treatment effects on test scores and essay quality, separately for the full sample (Panel A) and by how students used AI (Panel B). Overall (TOT) is the treatment-on-the-treated estimate, instrumenting actual ChatGPT use with random assignment to the AI-allowed group. Treated students' conversation logs are classified into four mutually exclusive categories (see Appendix B.6): *Automation*, if the AI did the work *for* the student; *Augmentation*, if the AI worked *with* the student (e.g., explaining concepts or giving feedback on the student's own writing); *Mixed*, if both occurred within the same conversation; and *Other*, if the conversation was off-topic. Panel B reports effects for two subgroups, each pooled with the control group as the comparison: *Automation users* (Automation + Mixed) and *Augmentation users* (Augmentation + Mixed). A student belongs to a subgroup if any of their conversations falls in it, so students with a Mixed conversation—or with separate Automation and Augmentation conversations—appear in both rows; the two subgroups are therefore not mutually exclusive, and the coefficients need not average to the Panel A overall effect. Treated students whose conversations were classified as Other, or who did not use ChatGPT at all, appear only in Panel A. All specifications use controls selected by double-lasso on the full sample (Belloni et al., 2014), with strata fixed effects. 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.

Andre, P., Haaland, I., Roth, C., Wiederholt, M., and Wohlfart, J. (2026). Narratives about the macroeconomy. *Review of Economic Studies*. Forthcoming.

Ba, H., Zhang, L., and Yi, Z. (2024). Enhancing clinical skills in pediatric trainees: A comparative study of ChatGPT-assisted and traditional teaching methods. *BMC Medical Education*, 24:558.

Baird, M., Carpanelli, M., Xu, B., and Xu, K. (2026). Firms' GitHub Copilot adoption and labor market outcomes for software engineers. *Contemporary Economic Policy*.

Bassner, P., Lenk-Ostendorf, B., Beinstingel, R., Wasner, T., and Krusche, S. (2026). Less stress, better scores, same learning: The dissociation of performance and learning in AI-supported programming education. *Computers and Education: Artificial Intelligence*, 10:100537.

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. Correction at https://doi.org/10.1073/pnas.2518204122.

Becker, J., Rush, N., Barnes, B., 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.

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., and Wadman, K. (2025). How people use ChatGPT. Working Paper 34255, National Bureau of Economic Research.

Chi, M. T. H. and Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4):219–243.

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.

Choi, J. H., Monahan, A. B., and Schwarcz, D. (2024). Lawyering in the age of artificial intelligence. *Minnesota Law Review*, 109(1):147–218.

Chung, A. T.-H., Zhang, B., Kung, L.-C., Bastani, H., and Bastani, O. (2026). Effective personalized AI tutors via LLM-Guided reinforcement learning. SSRN Working Paper 6423358, University of Pennsylvania.

Contractor, Z. and Reyes, G. (2026). 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.

Cruces, G., Fernández Meijide, D., Galiani, S., Gálvez, R. H., and Lombardi, M. (2026). Does generative AI narrow education-based productivity gaps? Evidence from a randomized experiment. Technical Report 34851, National Bureau of Economic Research.

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*.

Dai, X., Wen, Z., Jiang, J., Liu, H., and Zhang, Y. (2025). How students use AI feedback matters: Experimental evidence on physics achievement and autonomy.

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

Dell'Acqua, F., Ayoubi, C., Lifshitz, H., Sadun, R., Mollick, E., Mollick, L., Han, Y., Goldman, J., Nair, H., Taub, S., and Lakhani, K. R. (2025). The cybernetic teammate: A field experiment on generative AI reshaping teamwork and expertise. Technical Report 33641, National Bureau of Economic Research.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2026). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality. *Organization Science*.

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.

Doshi, A. R. and Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290.

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

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.

Fischer, M., Rau, H. A., and Rilke, R. M. (2025). AI tutoring enhances student learning without crowding out reading effort. IZA Discussion Paper 18338, IZA Institute of Labor Economics.

Gan, W., Ouyang, J., Li, H., Xue, Z., Zhang, Y., Dong, Q., Huang, J., Zheng, X., and Zhang, Y. (2024). Integrating ChatGPT in orthopedic education for medical undergraduates: Randomized controlled trial. *Journal of Medical Internet Research*, 26:e57037.

Goldsmith-Pinkham, P., Tan, C., and Zentefis, A. K. (2026). Human-AI collaboration in radiology: The case of pulmonary embolism. *arXiv preprint arXiv:2601.13379*.

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.

Hirabayashi, S., Jain, R., Jurković, N., and Wu, G. (2024). Harvard undergraduate survey on generative AI. *arXiv preprint arXiv:2406.00833*.

Hou, X., Xiao, B., Liu, H., and Mueller, S. (2026). The role of instructional guidance in generative AI-assisted learning: Empirical evidence from construction engineering education.

Huang, S., Wen, C., Bai, X., Li, S., Wang, S., Wang, X., and Yang, D. (2025). Exploring the application capability of ChatGPT as an instructor in skills education for dental medical students: Randomized controlled trial. *Journal of Medical Internet Research*, 27:e68538.

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 Weblogs and Social Media (ICWSM)*, volume 8, pages 216–225.

Jabarian, B. and Imas, A. (2025). Artificial writing and automated detection. NBER Working Paper 34223, National Bureau of Economic Research.

Jackson, C. K. and Mackevicius, C. L. (2024). What impacts can we expect from school spending policy? Evidence from evaluations in the U.S. *American Economic Journal: Applied Economics*, 16(1):412–446.

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

Jia, N., Luo, X., Fang, Z., and Liao, C. (2024). When and how artificial intelligence augments employee creativity. *Academy of Management Journal*, 67(1):5–32.

Kanazawa, K., Kawaguchi, D., Shigeoka, H., and Watanabe, Y. (2025). AI, skill, and productivity: The case of taxi drivers. *Management Science*.

Kavadella, A., Dias da Silva, M. A., Kaklamanos, E. G., Stamatopoulos, V., and Giannakopoulos, K. (2024). Evaluation of ChatGPT's real-life implementation in undergraduate dental education: Mixed methods study. *JMIR Medical Education*, 10:e51344.

Kazemitabaar, M., Chow, J., Ma, C. K. T., Ericson, B. J., Weintrop, D., and Grossman, T. (2023). Studying the effect of AI code generators on supporting novice learners in introductory programming. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM.

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, D., Mitrofanov, D., Wen, Q., and Xu, T. (2025). Generative AI can improve performance and engagement without harming learning. SSRN Working Paper 5929576.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4):241–253.

Kreijkes, P., Kewenig, V., Kuvalja, M., Lee, M., Vitello, S., Hofman, J. M., Sellen, A., Rintel, S., Goldstein, D. G., Rothschild, D. M., Tankelevitch, L., and Oates, T. (2026). Effects of LLM use and note-taking on reading comprehension and memory: A randomised experiment in secondary schools. *Computers & Education*, 243:105514.

Kumar, H., Rothschild, D. M., Goldstein, D. G., and Hofman, J. M. (2025). Math education with large language models: Peril or promise? In *Artificial Intelligence in Education (AIED 2025)*, Lecture Notes in Computer Science. Springer.

LearnLM Team (2024). LearnLM: Improving Gemini for Learning. Google DeepMind Technical Report.

LearnLM Team (2026). Teaching with Gemini: Measuring the impact of Guided Learning on student mathematics progress in Sierra Leone. Technical report, Google DeepMind and Fab AI.

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, G., Christian, B., Dumbalska, T., Bakker, M. A., and Dubey, R. (2026). AI assistance reduces persistence and hurts independent performance.

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. (2025). Introducing Pangram's plagiarism detection. Pangram Labs blog post.

Masrour, E., Emi, B. N., and Spero, M. (2025). DAMAGE: Detecting adversarially modified AI generated text. In *Proceedings of the 1st Workshop on Detecting AI Generated Content (GenAIDetect), COLING*, pages 71–86.

Meincke, L., Nave, G., and Terwiesch, C. (2025). ChatGPT decreases idea diversity in brainstorming. *Nature Human Behaviour*, 9(6):1107–1109.

Moon, K., Green, A. E., and Kushlev, K. (2025). Homogenizing effect of large language models (LLMs) on creative diversity: An empirical comparison of human and ChatGPT writing. *Computers in Human Behavior: Artificial Humans*, 6:100207.

Moon, K., Kushlev, K., Bank, A., Lira Luttges, B., Viskontas, I., Kaufman, J. C., Johnson, D. R., Duckworth, A., and Green, A. (2026). The creative link between words and ideas is weakening in the AI era. PsyArXiv preprint.

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.

Nickow, A., Oreopoulos, P., and Quan, V. (2024). The promise of tutoring for PreK–12 learning: A systematic review and meta-analysis of the experimental evidence. *American Educational Research Journal*, 61(1):74–107.

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.

Poulidis, S., Bastani, H., and Bastani, O. (2025). Self-regulated AI use hinders long-term learning. SSRN Working Paper 5604932, University of Pennsylvania.

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.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992. Association for Computational Linguistics.

Risko, E. F. and Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9):676–688.

Shen, J. H. and Tamkin, A. (2026). How AI impacts skill formation.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2020). MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 16857–16867.

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.

Strömberg, D., Lei, V., and Wu, Y. (2026). The generative AI learning penalty: Evidence from Chinese secondary education. CEPR Discussion Paper 21577, Centre for Economic Policy Research.

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, X., Qiao, L., Cheng, N., Liu, H., and Zhao, W. (2025). Enhancing self-regulated learning and learning experience in generative AI environments: The critical role of metacognitive support. *British Journal of Educational Technology*, 56(5):1842–1863.

Yakura, H., Lopez-Lopez, E., Brinkmann, L., Serna, I., Gupta, P., Soraperra, I., and Rahwan, I. (2024). 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

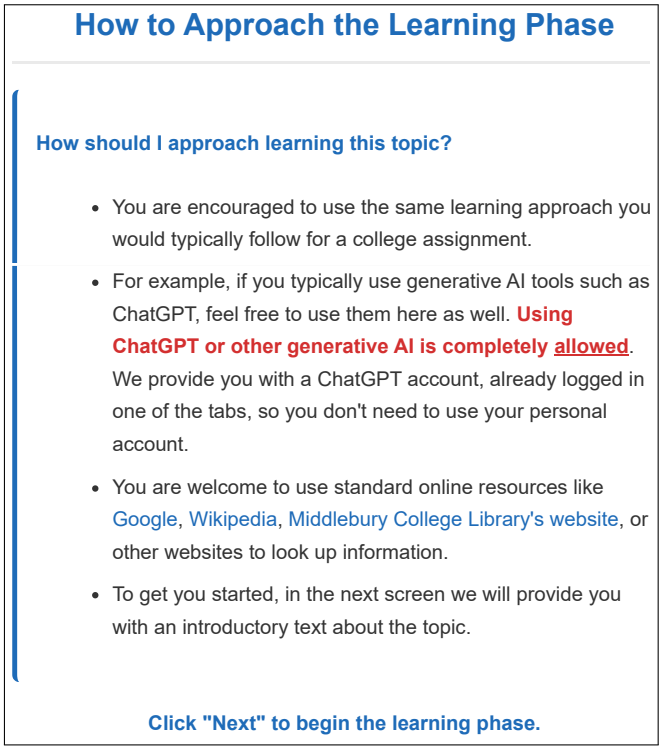
A Appendix Figures and Tables

Figure A1: Computer Lab Setup with Privacy Dividers

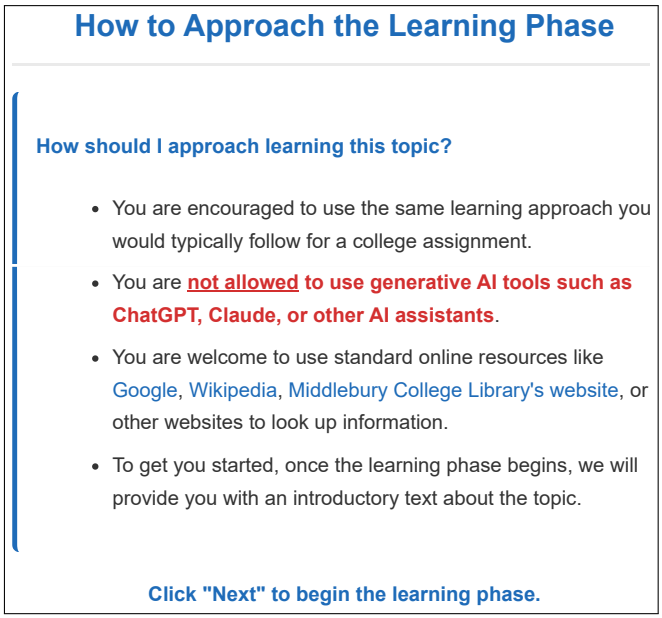
Notes: This figure shows the computer lab setup used during experimental sessions. Each workstation was equipped with privacy dividers to minimize distractions and prevent participants from viewing other screens.

Figure A2: Treatment Instructions Displayed to Participants

Panel A. AI-allowed condition



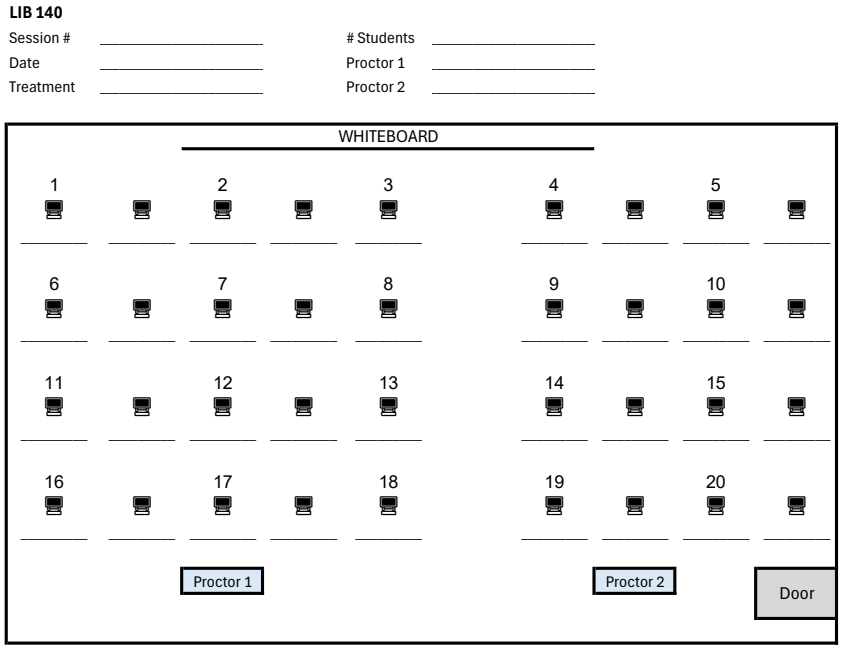
Panel B. AI-forbidden condition



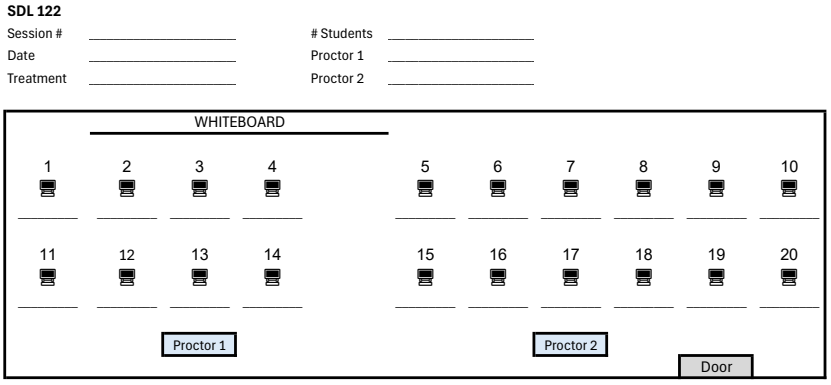
Notes: This figure shows the treatment-specific instructions displayed to participants in the survey interface.

Figure A3: Laboratory Seating Charts

Panel A. Library computer lab (LIB 140)

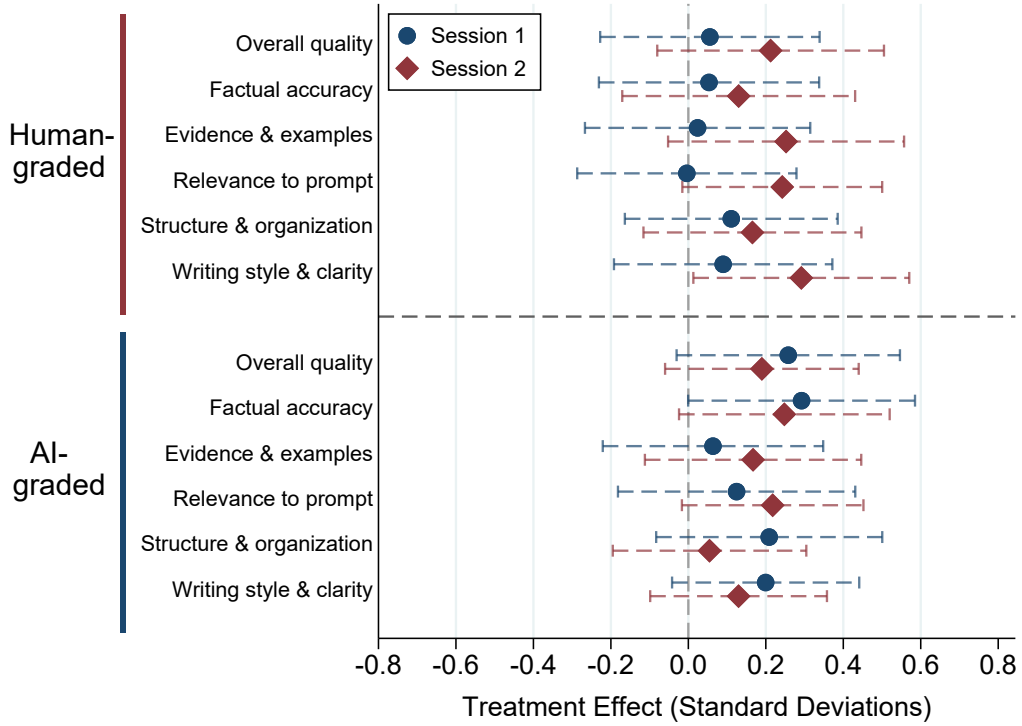


Panel B. Language Building computer lab (SDL 122)



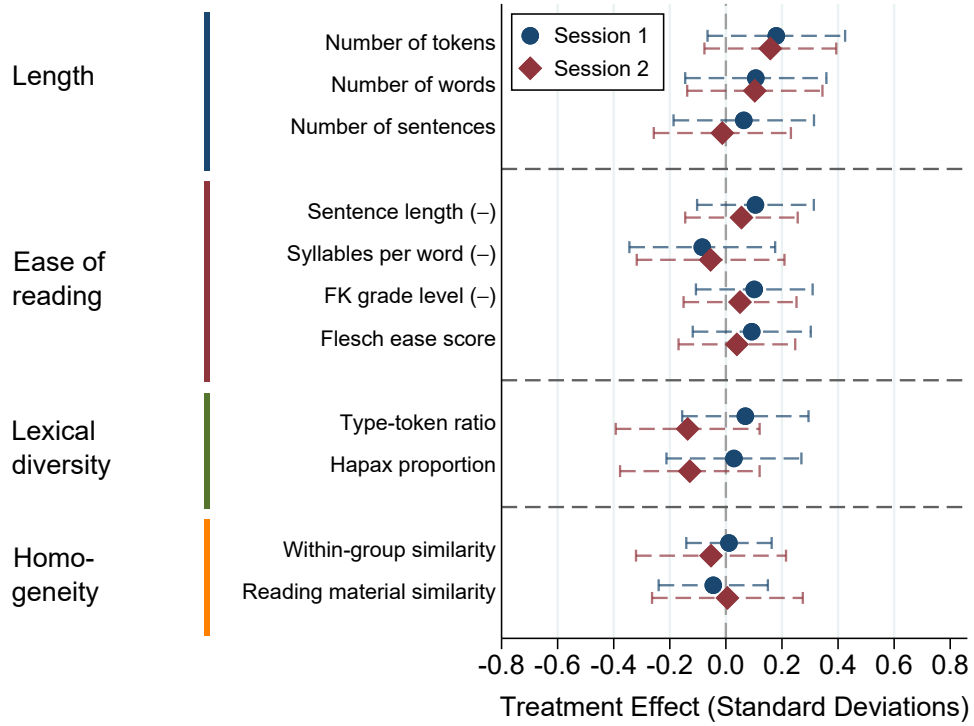
Notes: This figure shows the seating charts used by laboratory staff to monitor compliance during experimental sessions. Each numbered square represents a computer workstation. For each session time slot, one lab was randomly assigned to the AI-allowed condition and the other to the AI-forbidden condition. 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 helped prevent cross-contamination between experimental groups.

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



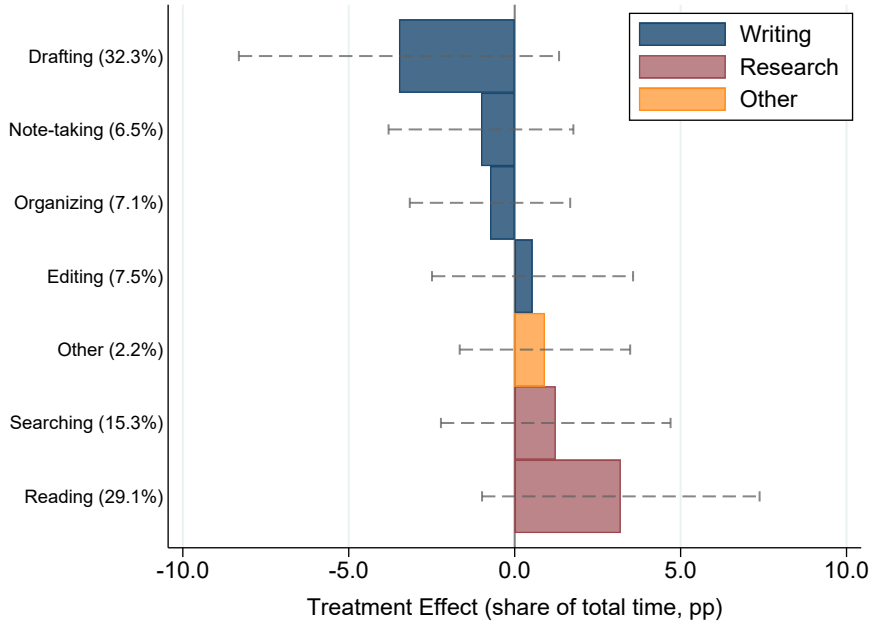
Notes: This figure presents treatment effects of AI access on individual essay quality dimensions, measured in standard deviations. The top panel shows human-graded dimensions: essays were evaluated by independent graders on five rubric dimensions plus an overall quality rating (factual accuracy, use of evidence and examples, relevance to the prompt, structure and organization, writing style and clarity, and overall quality). Human-graded regressions are at the student level, averaging each essay's scores across its graders. The bottom panel shows the same dimensions as scored by an AI grader (see Appendix B.4). Each dimension is scored on a scale from 0 to 10 and standardized to have mean zero and standard deviation one 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 based on heteroskedasticity-robust standard errors.

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 and similarity (within-group cosine similarity and reading material cosine similarity). Circles represent Session One effects; diamonds represent Session Two effects. Horizontal lines represent 95 percent confidence intervals based on heteroskedasticity-robust standard errors.

Figure A6: Mechanisms: Time Allocation Across Learning Activities

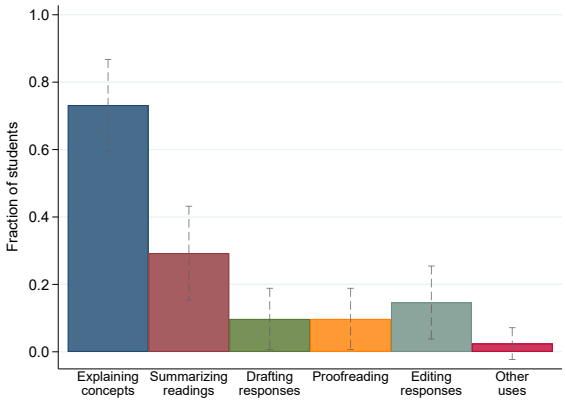


Notes: This figure presents treatment effects of AI access on self-reported time allocation across different learning activities, measured as each activity's share of the student's total self-reported learning time (in percentage points). Students reported how many minutes they spent on each activity during the 35-minute learning phase; each activity's share is its minutes divided by the sum of minutes across all activities. 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, maroon 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 share.

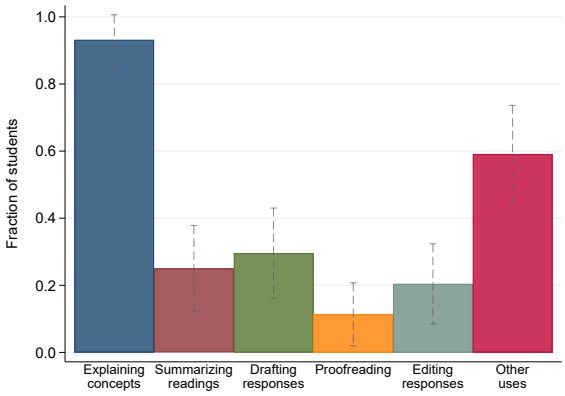
Figure A7: Types of Generative AI Use by Automation and Augmentation Users

Augmentation users

Panel A. Self-reported usage

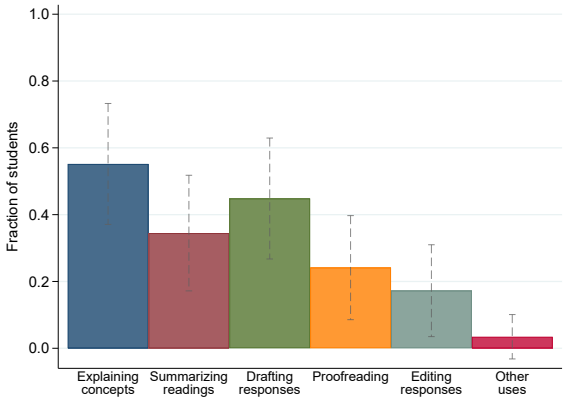


Panel B. Revealed usage (ChatGPT logs)

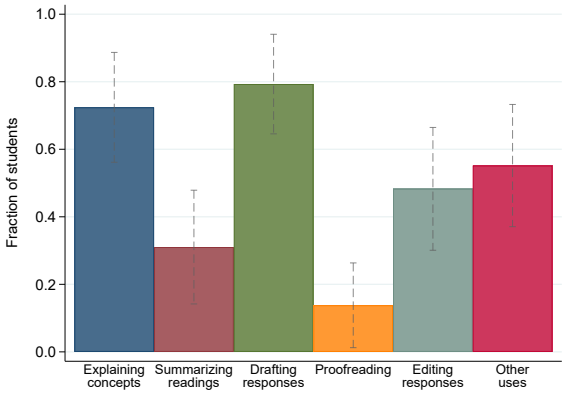


Automation users

Panel C. Self-reported usage



Panel D. Revealed usage (ChatGPT logs)

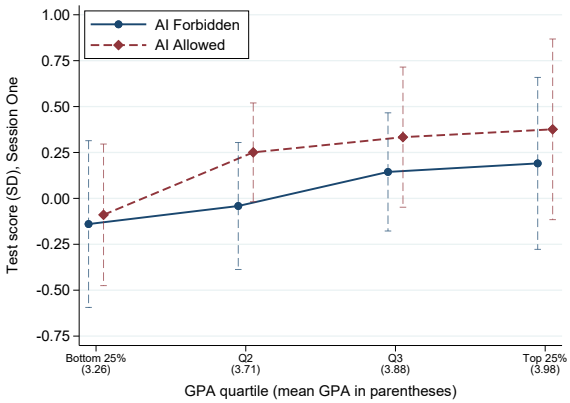


Notes: This figure replicates the analysis from Figure 3, separately for augmentation and automation users as classified by Claude Opus (see Appendix B.6). Panels A and C present self-reported usage types from the exit survey; participants could select multiple categories. Panels B and D show the fraction 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. Vertical bars denote 95 percent confidence intervals.

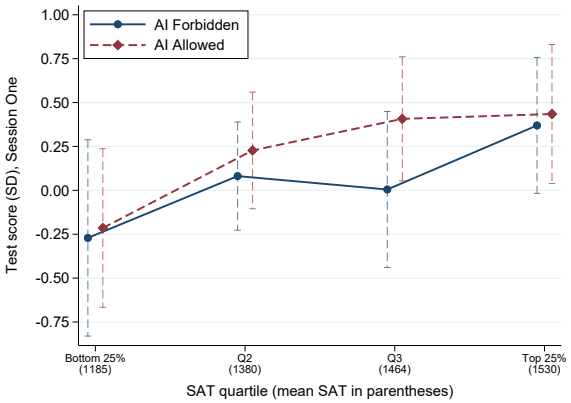
Figure A8: Treatment Effects on Test Scores by GPA and SAT Quartile

Session One

Panel A. By GPA quartile

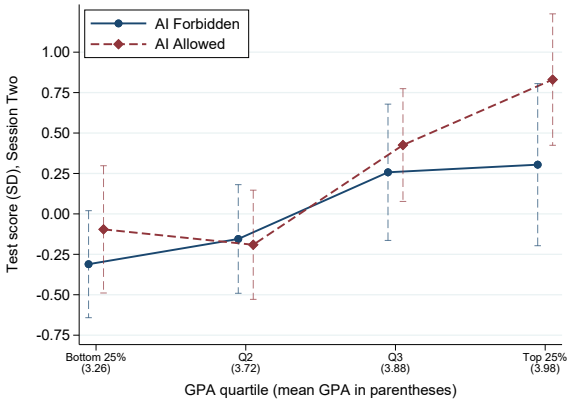


Panel B. By SAT quartile

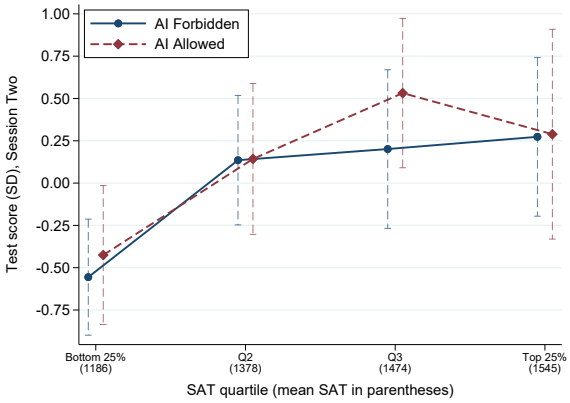


Session Two

Panel C. By GPA quartile



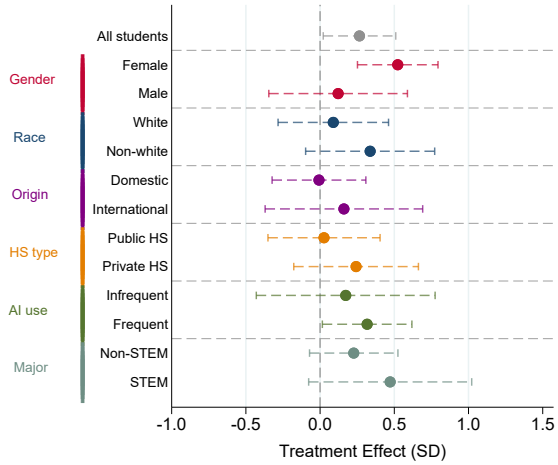
Panel D. By SAT quartile



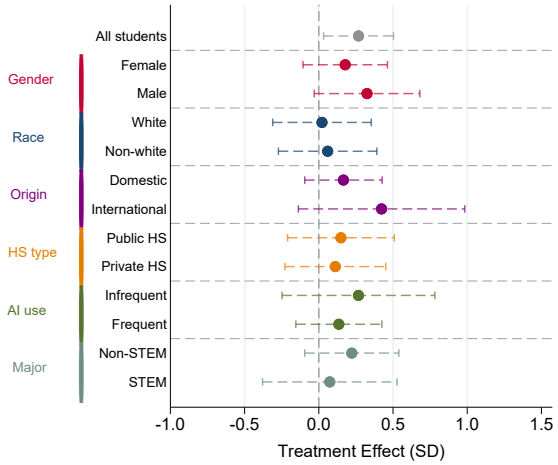
Notes: This figure shows mean test scores by treatment status across quartiles of GPA (Panels A and C) and SAT scores (Panels B and D). The SAT panels are restricted to students who reported an SAT score; quartiles of GPA include all attendees. Numbers in parentheses below each tick label are within-quartile means of the splitting variable. The four SAT quartile means (1185, 1380, 1464, 1530) correspond approximately to the 75th, 92nd, 97th, and 99th percentiles of the College Board user-percentile distribution for SAT test-takers. Test scores are standardized to have mean zero and standard deviation one in the control group. Vertical bars denote 95 percent confidence intervals.

Figure A9: Heterogeneity in Treatment Effects by Student Characteristics

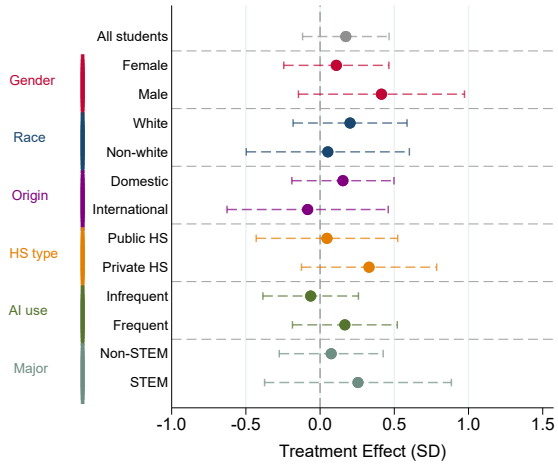
Panel A. Test score, Session One



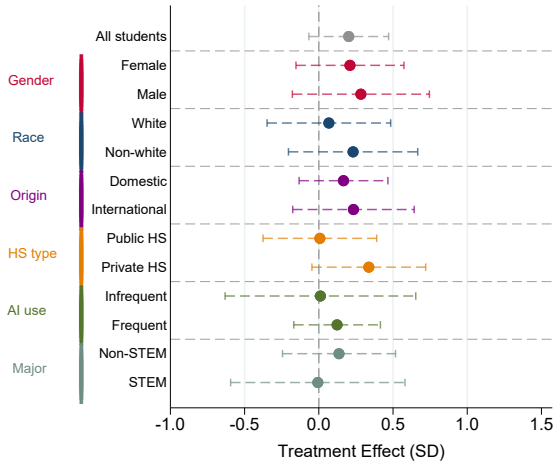
Panel B. Test score, Session Two



Panel C. Essay quality, Session One

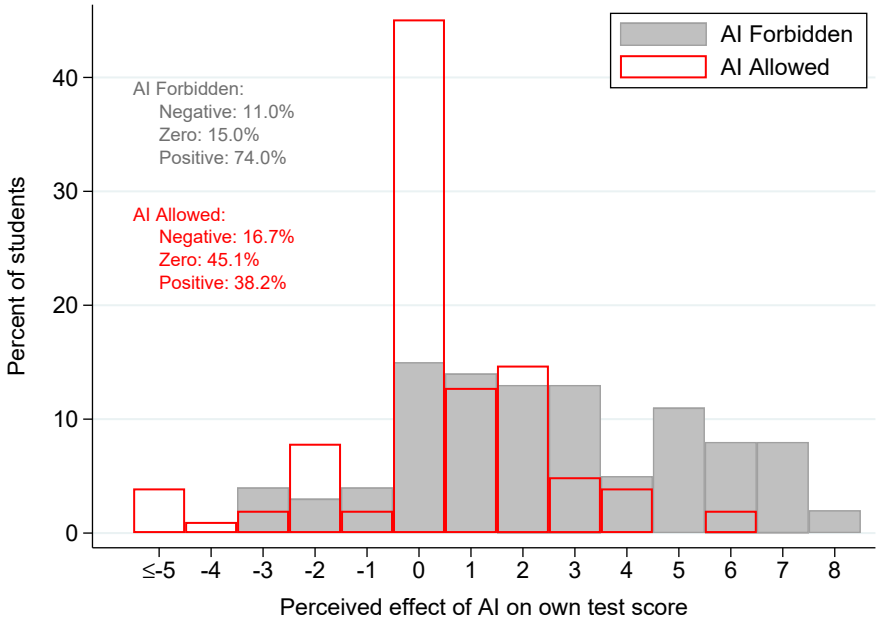


Panel D. Essay quality, Session Two



Notes: This figure presents heterogeneity in the intent-to-treat effect of AI access across student subgroups. Panels A and B show effects on test score performance in Sessions One and Two; Panels C and D show effects on overall essay quality (the average of the human-graded and AI-graded overall scores) in Sessions One and Two. Within each panel, the top point ("All students") reports the full-sample ITT, while subsequent points report subgroup-specific ITT estimates from separate regressions on each subgroup. All effects are expressed in standard deviations of the control-group outcome distribution. Each regression includes strata fixed effects and controls selected by double-lasso within the subgroup sample (Belloni et al., 2014), with heteroskedasticity-robust standard errors. Subgroups defined by median splits use the full-sample median; "AI use" splits students by frequency of AI use for academics. Horizontal dashed bars denote 95 percent confidence intervals.

Figure A10: Distribution of Perceived Treatment Effects on Own Test Score



Notes: This figure plots the distribution of students' perceived treatment effect of AI access on their own Session Two test score, separately for AI-forbidden (control) and AI-allowed (treated) students. The perceived effect is each student's believed test score with AI access minus their believed score without it, expressed in number of questions answered correctly (for treated students, the believed own score minus the believed counterfactual; for control students, the believed counterfactual minus the believed own score). Values below -5 are bottom-coded at -5. Bin width is 1 and bins are centered on integers. In-figure annotations report the share of each group reporting a negative, zero, or positive perceived effect.

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	0.138	0.113	0.098	0.031
	(0.104)	(0.110)	(0.110)	(0.109)
White	−0.218**	−0.065	−0.194*	0.051
	(0.107)	(0.112)	(0.114)	(0.138)
International student	0.166	0.093	0.169	0.134
	(0.115)	(0.127)	(0.150)	(0.146)
Private high school	−0.043	−0.083	−0.176	−0.208**
	(0.096)	(0.106)	(0.109)	(0.100)
Panel B. Academic background				
Natural Sciences major	−0.088	−0.167	−0.121	−0.125
	(0.100)	(0.106)	(0.127)	(0.129)
Social Sciences major	0.103	0.258**	0.088	0.274**
	(0.099)	(0.101)	(0.121)	(0.125)
High GPA (above median)	−0.142	−0.101	−0.004	0.010
	(0.090)	(0.100)	(0.110)	(0.128)
Freshman	−0.112	−0.071	−0.209	−0.004
	(0.099)	(0.109)	(0.136)	(0.147)
Sophomore	0.082	0.119	−0.094	0.034
	(0.105)	(0.125)	(0.128)	(0.174)
Junior	−0.066	−0.016	−0.125	−0.085
	(0.127)	(0.133)	(0.142)	(0.155)
Panel C. AI experience and beliefs				
Frequent AI user (above median)	0.334***	0.359***	0.241*	0.349**
	(0.108)	(0.116)	(0.145)	(0.164)
Early AI adopter (before Fall 2024)	0.176*	0.161	−0.012	−0.071
	(0.094)	(0.103)	(0.111)	(0.131)
High AI proficiency (above median)	0.231**	0.264**	0.093	0.151
	(0.092)	(0.103)	(0.112)	(0.146)
High perceived AI effect (above median)	0.001	−0.185	0.070	−0.081
	(0.110)	(0.134)	(0.113)	(0.128)
N (Students)	107	102	107	102

Notes: This table reports the association between student characteristics and AI use among treated 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 above-median frequency of AI use during the semester (occasional or more frequent). Early AI 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 include session (strata) fixed effects. Heteroskedasticity-robust standard errors in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Table A2: 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	0.139	0.216
		(0.211)	(0.328)		(0.197)	(0.303)
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	−0.134	0.325	0.664
		(0.245)	(0.352)		(0.356)	(0.791)
N	34	69	69	33	66	66
Panel C. Carbon capture						
Self−assessed	4.839	0.185	0.284	4.000	0.113	0.171
		(0.409)	(0.632)		(0.375)	(0.574)
Frac. correct	0.516	0.107*	0.164*	0.413	0.012	0.020
		(0.062)	(0.098)		(0.034)	(0.055)
Test score (SD)	−0.187	0.420*	0.647*	−0.433	0.064	0.107
		(0.245)	(0.387)		(0.180)	(0.289)
N	31	65	65	31	63	63

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. Heteroskedasticity-robust standard errors in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Table A3: Robustness of the Main Results to Sample Restrictions

Outcome	Session One			Session Two		
	Control mean (1)	ITT (2)	TOT (3)	Control mean (4)	ITT (5)	TOT (6)
Panel A. Excluding failed attention check						
Self−assessed	4.750	0.304	0.492	3.823	0.409*	0.641*
		(0.238)	(0.386)		(0.243)	(0.384)
Fraction correct	0.550	0.097**	0.155**	0.508	0.055*	0.086*
		(0.039)	(0.062)		(0.030)	(0.048)
Test score (SD)	−0.053	0.381**	0.611**	0.068	0.287*	0.452*
		(0.153)	(0.246)		(0.158)	(0.252)
N	64	131	131	62	127	127
Panel B. Excluding failed comprehension						
Self−assessed	4.819	0.067	0.101	3.793	0.112	0.165
		(0.223)	(0.337)		(0.236)	(0.346)
Fraction correct	0.583	0.062*	0.092*	0.500	0.066***	0.100**
		(0.037)	(0.055)		(0.025)	(0.039)
Test score (SD)	0.078	0.243*	0.362*	0.026	0.350***	0.527**
		(0.145)	(0.218)		(0.134)	(0.205)
N	83	167	167	82	161	161
Panel C. Excluding studied between sessions						
Self−assessed	4.786	−0.026	−0.038	3.832	0.128	0.188
		(0.191)	(0.283)		(0.201)	(0.293)
Fraction correct	0.563	0.067**	0.100**	0.500	0.044**	0.065**
		(0.032)	(0.048)		(0.021)	(0.033)
Test score (SD)	−0.001	0.263**	0.393**	0.026	0.230**	0.343**
		(0.125)	(0.188)		(0.112)	(0.171)
N	103	210	210	101	203	203
Panel D. Excluding looked-up topic						
Self−assessed	4.763	−0.027	−0.040	3.811	0.175	0.255
		(0.193)	(0.285)		(0.205)	(0.297)
Fraction correct	0.573	0.056*	0.084*	0.504	0.044**	0.065*
		(0.033)	(0.050)		(0.022)	(0.033)
Test score (SD)	0.038	0.219*	0.330*	0.048	0.231**	0.342*
		(0.131)	(0.199)		(0.115)	(0.174)
N	97	201	201	95	193	193

Notes: This table replicates the main test-score results (Table 4) under alternative sample restrictions. Each panel restricts the sample: *Excluding failed attention check* drops participants who failed the attention check; *Excluding failed comprehension* drops those who answered fewer than five of five comprehension questions correctly; *Excluding studied between sessions* drops those who reported studying the topic between sessions; *Excluding looked-up topic* drops those who reported looking up the topic between sessions. 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. Heteroskedasticity-robust standard errors in parentheses. The *N* row reports the test-score sample size for each panel. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Table A4: The Distributional Impact of Generative AI on Test Scores

	Session One			Session Two		
Outcome	Control mean (1)	ITT (2)	TOT (3)	Control mean (4)	ITT (5)	TOT (6)
Above 20% correct	0.990	−0.010	−0.014	0.951	0.034	0.050
		(0.016)	(0.024)		(0.025)	(0.036)
Above 40% correct	0.827	0.080*	0.118	0.804	0.005	0.007
		(0.048)	(0.072)		(0.051)	(0.075)
Above 60% correct	0.577	0.096	0.142	0.353	0.122*	0.181**
		(0.064)	(0.094)		(0.062)	(0.092)
Above 80% correct	0.317	0.019	0.029	0.098	0.044	0.064
		(0.061)	(0.093)		(0.042)	(0.061)
100% correct	0.106	0.037	0.055	0.010	−0.009	−0.013
		(0.043)	(0.064)		(0.009)	(0.013)
N	104	211	211	102	204	204

Notes: This table reports treatment effects of AI access on the probability of scoring at or above various thresholds in Session One and Session Two. 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 A5: Effects of 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	0.029	0.009	0.013
		(0.009)	(0.014)		(0.017)	(0.025)
Tokens	202.981	17.857	27.559	150.869	10.756	15.808
		(12.422)	(19.193)		(8.140)	(11.990)
Words	347.875	16.769	25.025	276.424	12.993	19.380
		(20.193)	(30.143)		(15.499)	(23.154)
Sentences	16.433	0.505	0.770	12.545	−0.070	−0.103
		(1.012)	(1.540)		(0.693)	(1.019)
Panel B. Readability						
Sentence length	27.119	−4.655	−6.947	25.565	−1.816	−2.677
		(4.688)	(7.009)		(3.346)	(4.935)
Syllables/word	1.775	0.010	0.015	1.713	0.007	0.010
		(0.016)	(0.024)		(0.017)	(0.025)
FK grade level	15.936	−1.736	−2.591	14.597	−0.626	−0.923
		(1.820)	(2.723)		(1.276)	(1.882)
Flesch ease	29.108	4.155	6.201	35.946	1.255	1.850
		(4.837)	(7.239)		(3.392)	(4.999)
Panel C. Lexical diversity						
TTR	0.723	0.006	0.010	0.717	−0.014	−0.020
		(0.010)	(0.016)		(0.011)	(0.017)
Hapax prop.	0.579	0.003	0.005	0.565	−0.017	−0.025
		(0.014)	(0.021)		(0.017)	(0.024)
Panel D. Homogeneity and similarity						
Cosine sim.	0.774	0.001	0.001	0.762	−0.005	−0.007
		(0.007)	(0.010)		(0.013)	(0.019)
Reading material sim.	0.746	−0.004	−0.006	0.704	0.001	0.001
		(0.009)	(0.014)		(0.013)	(0.019)
N	104	209	209	99	199	199

Notes: This table reports treatment effects on individual essay characteristics. Each row reports the effect of being assigned to the AI-allowed group on the indicated outcome. Essay is empty is an indicator equal to one if the essay contains no text. Tokens is the number of content tokens in the essay (after removing stopwords, numbers, and punctuation). Words is the total number of words. Sentences is the total number of sentences. Sentence length is the mean number of words per sentence. Syllables/word is the mean number of syllables per word. FK grade level is the Flesch-Kincaid Grade Level. 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. Reading material sim. is the average cosine similarity between each essay and the provided reading material. 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 A6: Effects of AI Access on Essay Quality, by Dimension and Grader Type

	Session One			Session Two		
Outcome	Control mean (1)	ITT (2)	TOT (3)	Control mean (4)	ITT (5)	TOT (6)
Panel A. Human graders						
Accuracy	7.087	0.064	0.095	6.119	0.178	0.262
		(0.173)	(0.258)		(0.210)	(0.313)
Evidence	5.950	0.035	0.053	4.674	0.387	0.569
		(0.220)	(0.329)		(0.238)	(0.360)
Relevance	7.192	−0.006	−0.008	6.383	0.279	0.415
		(0.207)	(0.308)		(0.217)	(0.327)
Organization	6.235	0.162	0.242	5.548	0.261	0.385
		(0.205)	(0.306)		(0.227)	(0.339)
Writing style	6.401	0.121	0.180	5.580	0.418**	0.616**
		(0.194)	(0.289)		(0.204)	(0.311)
N	104	210	210	97	197	197
Panel B. AI graders						
Accuracy	6.404	0.413*	0.630*	4.222	0.466**	0.689**
		(0.212)	(0.324)		(0.232)	(0.349)
Evidence	6.337	0.117	0.175	3.465	0.264	0.395
		(0.268)	(0.399)		(0.225)	(0.339)
Relevance	6.721	0.211	0.318	5.687	0.449*	0.656*
		(0.264)	(0.397)		(0.235)	(0.348)
Organization	5.625	0.353	0.537	5.020	0.124	0.182
		(0.252)	(0.384)		(0.245)	(0.361)
Writing style	6.548	0.352	0.535	5.980	0.246	0.370
		(0.216)	(0.330)		(0.198)	(0.301)
N	104	209	209	99	198	198

Notes: This table reports treatment effects on essay quality dimensions, separately for human and AI graders. Each row reports the effect of being assigned to the AI-allowed group on the indicated essay quality dimension. Each dimension is scored on a 0–10 scale. Panel A reports results from human graders (Prolific workers, 3–4 graders per essay); Panel B reports results from AI grading (Claude Opus, one score per essay). Both sets of regressions are at the student level—human grades are averaged across each essay's 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). Intent-to-treat (ITT) estimates report the effect of being assigned to the AI-allowed group; treatment-on-the-treated (TOT) estimates instrument ChatGPT use with random 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 A7: 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)
Treated ×						
Female	0.458*	−0.136	−0.151	−0.115	0.014	−0.022
	(0.259)	(0.338)	(0.349)	(0.259)	(0.289)	(0.295)
International	0.331	−0.159	−0.221	0.456	0.289	0.342
	(0.295)	(0.384)	(0.388)	(0.287)	(0.288)	(0.295)
Nonwhite	0.257	−0.063	−0.009	0.195	0.140	0.201
	(0.263)	(0.300)	(0.305)	(0.252)	(0.289)	(0.295)
Public HS	−0.314	−0.346	−0.190	−0.205	−0.475*	−0.508*
	(0.263)	(0.308)	(0.314)	(0.240)	(0.279)	(0.281)
STEM	0.188	0.220	0.244	−0.150	0.134	0.090
	(0.259)	(0.322)	(0.318)	(0.259)	(0.291)	(0.303)
Frequent AI user	0.216	0.343	0.291	0.040	−0.154	−0.271
	(0.322)	(0.308)	(0.307)	(0.275)	(0.318)	(0.325)
Early AI adopter	0.006	0.645**	0.561*	0.150	0.004	0.001
	(0.261)	(0.288)	(0.293)	(0.240)	(0.280)	(0.282)
High AI proficiency	−0.139	0.210	0.033	−0.034	0.063	0.032
	(0.282)	(0.317)	(0.329)	(0.253)	(0.300)	(0.299)
N	211	209	209	204	197	197

Notes: This table reports heterogeneity of treatment effects by student characteristics. Each row reports 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. Heteroskedasticity-robust standard errors in parentheses. Subgroups defined by median splits use the full-sample median. The reported N is the maximum analysis sample; the prior-AI-experience rows (frequency of AI use, early adoption, and AI proficiency) are estimated on up to seven fewer students who did not report these characteristics, so the effective sample size varies by row. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Table A8: Treatment Effects on Belief Accuracy

Outcome	Control mean (1)	ITT (2)	TOT (3)
Perceived AI gain on self (pp)	25.200	−21.403***	−30.945***
		(3.454)	(5.379)
Perceived AI gain on others (pp)	14.060	−3.497	−5.159
		(3.059)	(4.509)
Absolute error, self (vs actual)	27.159	−11.931***	−17.342***
		(2.479)	(3.710)
Absolute error, others (vs actual)	20.956	−3.398	−4.959
		(2.157)	(3.184)
N	100	201	201

Notes: This table reports treatment effects of AI access on belief accuracy. The sample is students who attended Session Two. Outcomes are in percentage points: "Perceived AI gain on self" is the signed predicted treatment effect on the student's own Session Two test score; "Perceived AI gain on others" is the predicted effect on others' scores; the "Absolute error" rows report |perceived gain − actual ITT|, where the actual ITT is the Session Two estimate from Table 4. Column (1) reports the control-group mean of each outcome; columns (2) and (3) report ITT and TOT estimates. All specifications include controls selected through a double-lasso procedure (Belloni et al., 2014) with strata fixed effects. Heteroskedasticity-robust standard errors in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

B Empirical Appendix

B.1 Reading Materials

We wrote original texts on each of the three topics, designing them so that participants faced similar cognitive demands across topics. Each text follows a parallel structure, beginning with an overview of the technology, followed by its historical development, technical underpinnings, core applications, and limitations. The texts are information-rich but self-contained, requiring no prior knowledge of the topic. The full texts are available in the Supplementary Materials.

The reading materials are 1,253–1,399 words long, or about 5.0–5.6 minutes of reading at an average silent reading speed of 250 words per minute for college students (Carver, 1992; Brysbaert, 2019). They are closely matched across several readability metrics (Appendix Table B1). Flesch-Kincaid grade levels of 15.1–15.9 place them at a late-undergraduate reading level. Measures of lexical and syntactic complexity, including sentence and word length, also indicate a consistent level of difficulty 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

Notes: This table presents readability metrics for the three reading materials. Estimated reading times assume an average silent reading speed of 250 words per minute for college students (Carver, 1992; Brysbaert, 2019).

B.2 Writing Assessment Prompts

This appendix presents the essay prompts we developed for each of the three topics. For each topic, we designed two complementary prompts of comparable difficulty. Each prompt requires participants to analyze relationships between concepts, evaluate the relative importance of different factors, and support their analysis with specific examples. Each

participant is randomly assigned to receive one prompt in Session One and the other in Session Two.

Blockchain Technology

Prompt 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."

Prompt 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

Prompt 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."

Prompt 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

Prompt 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."

Prompt 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 recruit independent graders through Prolific to evaluate participants' analytical essays, restricting eligibility to individuals holding a master's or PhD degree. Graders are blind to treatment condition, participant identity, and all other experimental variables. Each grader evaluates five essays and is compensated with a $15 fixed payment plus eligibility for a $10 lottery bonus. To incentivize consistency, we tell graders that their chances of winning the bonus rise if their scores resemble those of other graders evaluating the same essays.

Before grading, each grader completes a comprehension check verifying their understanding of the task requirements and reviews two example essays with suggested scores and detailed rationales. The examples are 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 can calibrate their standards. Graders are 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 evaluate each essay on five rubric dimensions plus an overall quality rating, each on a 0–10 scale. *Accuracy of Content* measures the factual correctness 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 report 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 is evaluated by three to four independent graders. We use the overall quality rating as our benchmark measure of essay quality and report results for each dimension 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 4.8 to grade each essay on the same five rubric dimensions plus the overall quality rating, using the same 0–10 scale as the Prolific graders.

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. We constrain the model's output to a JSON schema: a brief justification followed by an integer score on the 0–10 scale. Requiring the justification first lets the score reflect the model's stated reasoning.

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 dimensions and their scales are as follows:

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 with introduction, comparison, context, and conclusion; 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; reads more like notes or a summary than a full essay. 7-8: Mostly clear and readable with minor awkward phrasing; lacks some polish or precision in language but avoids major issues. 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 with suggested scores for the relevant dimension. After presenting the calibration material, the prompt provides the student's assigned topic, the exact prompt the student was shown in that session, and the essay text. Students saw a different prompt in each of the two sessions, so grading against the session-specific prompt ensures the relevance dimension reflects the task the student actually faced. Empty essays, or those with fewer than 20 characters, are skipped and recorded as missing without querying the model.

B.5 SAT Score Imputation

We lack admissions test scores (SAT or ACT) for approximately 26 percent of participants (Middlebury College has test-optional admissions). To include these students in our regressions, we impute SAT scores for students with missing scores using baseline covariates and assign them to decile bins alongside students with observed scores.

We estimate an OLS regression of SAT scores on baseline characteristics among the 189 students with observed scores. The predictors are college GPA, high school GPA, age, gender, race and ethnicity indicators (Black, Latino, Asian, with white as the omitted

category), international student status, public and private high school indicators, academic field indicators (natural sciences, social sciences, humanities and arts), weekly study hours, whether the student has declared a major, cohort fixed effects, baseline self-assessment of topic knowledge, and baseline number of test questions answered correctly (out of five).

We include a missing indicator (equal to one for imputed observations) 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.

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 classified using an LLM (Anthropic's Claude Opus 4).

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 LLM 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. Across all 435 student prompts, the distribution is: Explaining concepts (44 percent), Other (26 percent), Drafting responses (11 percent), Editing responses (10 percent), Summarizing readings (7 percent), and Proofreading (2 percent).

The exact system prompt provided to the LLM is:

```
You are classifying student prompts sent to ChatGPT during a learning experiment.
Students were given a reading passage on a science topic (Blockchain, CRISPR, or Carbon
Capture) and asked to write an essay.  Some students had access to ChatGPT during this
process.

Classify each student prompt into exactly ONE of these categories:

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
6.  Other -- Anything that does not fit the above (e.g., off-topic chat, testing the
tool)

Respond with ONLY the category name, exactly as written above.  Nothing else.
```

To illustrate, Appendix 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	"rewrite this to make the argument clearer"
	"rewrite the whole thing now"
Proofreading	"use appropriate grammer and capitalization"
Other	"keyboard shortcuts to copy and paste"
	"i think that policy challenges is most difficult what with trump in office and everything"

Notes: This table presents example ChatGPT prompts illustrating the classification categories. 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 four 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); (3) *Mixed*, where the conversation contains substantial elements of both; and (4) *Other*, where the conversation is off-topic, unrelated to the academic task, or consists of testing or experimenting with ChatGPT without engaging with the reading or essay.

For each conversation, we provide the LLM with the full exchange between the student and ChatGPT, along with descriptions of each category. Across all ChatGPT conversations, the distribution is: Augmentation (49 percent), Automation (31 percent), Mixed (5 percent), and Other (15 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.

The exact system prompt provided to Claude Opus 4 is:

```
You are classifying a full ChatGPT conversation from a learning experiment.  Students
were given a reading passage on a science topic (Blockchain, CRISPR, or Carbon Capture)
and asked to write an essay.  Some students had access to ChatGPT during this process.

Classify the overall conversation into exactly ONE of these categories:

1.  Automation -- The AI is doing the work FOR the student.  The student asks ChatGPT
to generate essay text, produce draft paragraphs, write responses, or create content
that the student could paste directly into their essay.  The student outsources the
writing/thinking to the AI.

2.  Augmentation -- The AI is working WITH the student.  The student uses ChatGPT to
understand concepts, get explanations, ask clarifying questions, get feedback on their
own writing, or check their work.  The student retains ownership of the thinking and
writing process.

3.  Mixed -- The conversation contains substantial elements of both automation and
augmentation, making it hard to assign a single label.

4.  Other -- The conversation is off-topic, unrelated to the academic task, or consists
of testing/experimenting with ChatGPT without engaging with the essay or reading
material.

Respond with ONLY the category name (Automation, Augmentation, Mixed, or Other).
Nothing else.
```

Figure B1: Automation Example: Student Delegates Writing to ChatGPT

Student

Write a response paper of at least 500 words… Analyze two application areas where blockchain has the strongest potential for positive impact.

ChatGPT

[Produces a complete 500-word essay on supply chain management and healthcare.]

Student

[Pastes a lightly edited version of the essay.] How many words does it have?

ChatGPT

567 words.

Notes: This figure shows an excerpt of a ChatGPT conversation classified as automation. 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 a 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: This figure shows an excerpt of a ChatGPT conversation classified as augmentation. 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 or Mixed, and as an *augmentation user* if any was labeled as Augmentation or Mixed. These groups are neither mutually exclusive nor collectively exhaustive: students with both types of conversations appear in both subsamples, and the "Other" category is excluded from the heterogeneity analysis. 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 experimental papers on generative AI and learning outcomes published or circulated between 2023 and 2026. We required that each study satisfy seven 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; (5) a clean no-AI control, rather than an active comparison such as human tutoring or structured hints; (6) a total sample size of at least 50 participants; and (7) a sample composed of students, rather than working professionals or general online participants.

Appendix Table B3 summarizes the 13 included studies; we describe each below.

Table B3: Included Studies in Literature Comparison

Study	Domain	N	Design	Outcome	Estimates
Ba et al. (2024)	Medicine	77	RCT	Unassisted exam	1
Bassner et al. (2026)	Coding	275	RCT	Unassisted post-test	2
Bastani et al. (2025)	Math	839	Cluster RCT	Unassisted exam	2
Dai et al. (2025)	Physics	349	RCT (×2)	Unassisted exam	3
De Simone et al. (2025)	Language	759	RCT	Unassisted post-test	2
Fischer et al. (2025)	Economics	334	Lab RCT	Unassisted exam	2
Gan et al. (2024)	Medicine	110	RCT	Unassisted exam	1
Hou et al. (2026)	Engineering	95	RCT	Unassisted post-test	2
Huang et al. (2025)	Medicine	187	RCT	Unassisted skill test	1
Kavadella et al. (2024)	Medicine	77	RCT	Unassisted exam	1
Kazemitabaar et al. (2023)	Coding	69	RCT	Post-test + 1-wk retention	2
LearnLM Team (2026)	Math	1,763	Cluster RCT	Unassisted exam	1
Lehmann et al. (2025)	Coding	176	Lab RCT (×2)	Unassisted post-test	2

Notes: This table lists the studies included in our literature comparison. 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.

Ba et al. (2024) randomized 77 medical interns rotating through pediatric cardiology at Sun Yat-sen University to two weeks of ChatGPT-assisted instruction or standard bedside teaching with identical cases and instructors. Scores on a closed-book knowledge exam were statistically indistinguishable (−0.07 SD), with both groups near the test ceiling. We computed the effect size and standard error from reported means and standard deviations.

Bassner et al. (2026) randomized 452 introductory-programming students at TU Munich to a 90-minute Java exercise with a scaffolded LLM tutor (Iris), unrestricted ChatGPT, or no AI (275 analyzed after exclusions). Both AI arms scored higher on the exercise

itself, but on the supervised, unassisted knowledge test neither arm learned more than the control: baseline-adjusted gains are −0.07 SD (Iris) and −0.01 SD (ChatGPT)—the paper's titular dissociation between performance and learning. We standardized the difference in pre–post gains by the pooled pre-test standard deviation.

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).

Dai et al. (2025) ran two randomized experiments with grade-10 physics students at a Chinese high school (Experiment 1: $N = 121$; Experiment 2: $N = 266$), providing LLM-generated feedback on regular homework over five weeks. In Experiment 1, treated students received recommended problems with AI heuristic hints; in Experiment 2, students could request AI help on demand while studying, either choosing the feedback type themselves or receiving the system's choice. Effects on the unassisted end-of-term physics exam are small and insignificant in all three AI arms relative to no-intervention controls (0.03–0.21 SD). Exam scores are standardized, so regression coefficients are effect sizes directly.

De Simone et al. (2025) studied 759 secondary school students in Nigeria randomly assigned to practice English language skills with an AI chatbot or a no-AI control. Students showed gains on unassisted post-tests in English skills (0.24 SD) and on the school's regular third-term curricular exam at the end of the intervention (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.

Fischer et al. (2025) randomized 334 university students in Berlin to study two microeconomics textbook excerpts with a GPT-4-based tutor grounded in the readings—available either throughout the session or only after ten minutes of independent reading—or with the textbook alone. On an incentivized, unassisted 25-question exam, unrestricted AI tutoring raised performance by 0.34 SD, while restricted access produced a statistically insignificant 0.13 SD. Test scores are standardized, so the reported coefficients are effect sizes directly.

Gan et al. (2024) randomized 110 third-year medical students at Jinan University to a one-week orthopedics review using ChatGPT-4 or conventional internet resources. The

ChatGPT group scored 0.40 SD higher on an unassisted 214-item orthopedics exam. We computed the effect size and standard error from reported means and standard deviations.

Hou et al. (2026) randomized 95 undergraduates with no prior construction-related coursework to review construction-engineering material with a generative-AI assistant—with or without a structured prompting framework—or with lecture slides only. On an unassisted post-test combining multiple-choice and open-ended items, guided AI use raised total scores by 0.86 SD, while unguided use produced a near-zero effect (0.09 SD). We computed effect sizes and standard errors from reported means and standard deviations.

Huang et al. (2025) randomized 187 dental students at Wuhan University to one week of operative-skills training with instructional videos plus ChatGPT-3.5 or the same videos alone. The ChatGPT group scored 0.67 SD higher on an unassisted virtual-reality skill assessment. We computed the effect size and standard error from reported means and standard deviations.

Kavadella et al. (2024) randomized 77 second-year dental students at European University Cyprus to complete a radiation-biology assignment using ChatGPT or conventional literature search. On an unannounced, unassisted ten-question exam, the ChatGPT group scored 0.52 SD higher. We computed the effect size and standard error from reported means and standard deviations.

Kazemitabaar et al. (2023) ran a 10-session randomized trial with 69 K-12 learners (ages 10–17, novice programmers) using OpenAI Codex to support Python authoring tasks. AI access produced large performance gains during training (1.67 SD) but no detectable difference on an immediate unassisted post-test (−0.05 SD; the baseline group scored slightly higher, 62.9 versus 61.3 percent) and only a marginal difference on a one-week retention test of code modification (0.41 SD). Standard errors were back-calculated from the reported N and standardized effect size.

LearnLM Team (2026) conducted a preregistered trial randomizing 48 mathematics classrooms (1,763 grade 7–8 students) in Sierra Leone to integrate Gemini's Guided Learning feature into half of weekly lessons for eight weeks or to continue standard instruction. The intent-to-treat effect on an unassisted, IRT-scaled endline mathematics assessment is 0.26 SD, with standard errors clustered at the classroom level.

Lehmann et al. (2025) ran two laboratory experiments with German 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 affected the treatment.

The random-effects grand mean in Figure 5 uses DerSimonian and Laird (1986) weights, accounting for between-study heterogeneity through a variance component estimated from study-level effects. The pooled estimate combines all 26 point estimates: 22 from the literature and 4 from this paper.

C Student Beliefs About AI and Learning

This section analyzes student responses to an open-ended question included in the Session Two survey: "In your opinion, how does generative AI (e.g., ChatGPT) affect student learning in college? Please explain your reasoning." Of 204 students who attended both sessions, all 204 provided a non-missing response. We analyze these responses in two ways: sentiment analysis to show the measure carries signal rather than noise, and causal-graph coding, following Andre et al. (2026), to characterize how students reason about AI and learning.

C.1 Validating the Open-Ended Response Measure

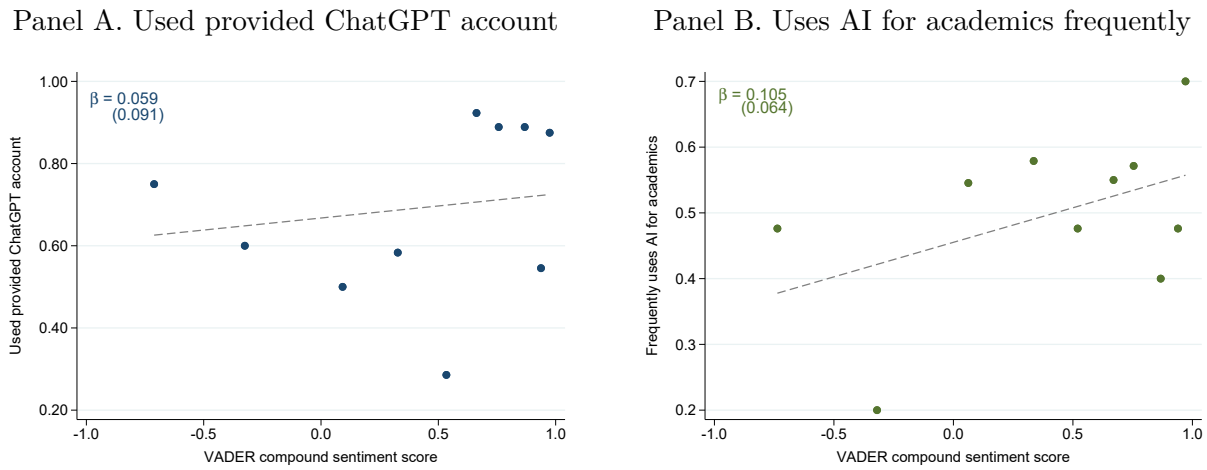
We validate our open-ended response measure using VADER (Valence Aware Dictionary and sEntiment Reasoner), a lexicon-based sentiment analysis tool (Hutto and Gilbert, 2014). VADER assigns each response a compound sentiment score ranging from −1 (most negative) to +1 (most positive). Appendix Figure C1 plots the distribution of these scores. Responses skew positive, with most students expressing net-positive sentiment about AI's effect on learning and a minority expressing negative sentiment.

Figure C1: Distribution of AI Sentiment Scores

Notes: This figure plots the distribution of compound sentiment scores across students who responded to the open-ended question about how generative AI affects student learning. Scores are computed using Hutto and Gilbert (2014)'s VADER algorithm and range from −1 (most negative) to +1 (most positive). The dashed vertical line marks the mean.

To assess whether the open-ended responses carry a relevant signal, we test whether their sentiment tracks students' actual AI use. Sentiment relates more strongly to students' AI use outside the experiment than to their take-up within it. Appendix Figure C2 presents binned scatterplots relating compound sentiment to two measures of AI use. Panel A relates sentiment to whether an AI-allowed student used AI during Session One. A regression of AI use on the sentiment score yields a positive but statistically insignificant coefficient ($\hat{\beta} = 0.059$, $p = 0.520$). Panel B shows that students who express more positive views about AI's effect on learning use AI somewhat more for academic work during the semester ($\hat{\beta} = 0.105$, $p = 0.103$). This pattern is consistent with the open-ended responses capturing a relevant signal.

Figure C2: 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 AI during Session One (from usage logs). Panel B shows an indicator for using AI for academic purposes frequently or very frequently during the semester, for all respondents. Each point represents the mean outcome for respondents within sentiment score 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.

C.2 Narratives as Causal Graphs

We characterize the content of the open-ended responses by coding each one as a causal graph, following Andre et al. (2026)'s analysis of how households and experts explain macroeconomic events. In their setting respondents explain why inflation rose, and the explanation is coded as a chain of cause-and-effect links running from a driving factor to

inflation. We adapt their method in one substantive way: their links are unsigned, because in the episode they study inflation only rose, whereas AI can raise or lower learning. Each link in our graphs therefore carries a sign. We represent each response as a signed causal graph from a source node, AI use, to a sink node, learning, passing through the mechanisms the student names. A link is positive if the student describes the mechanism as raising learning and negative if as lowering it.

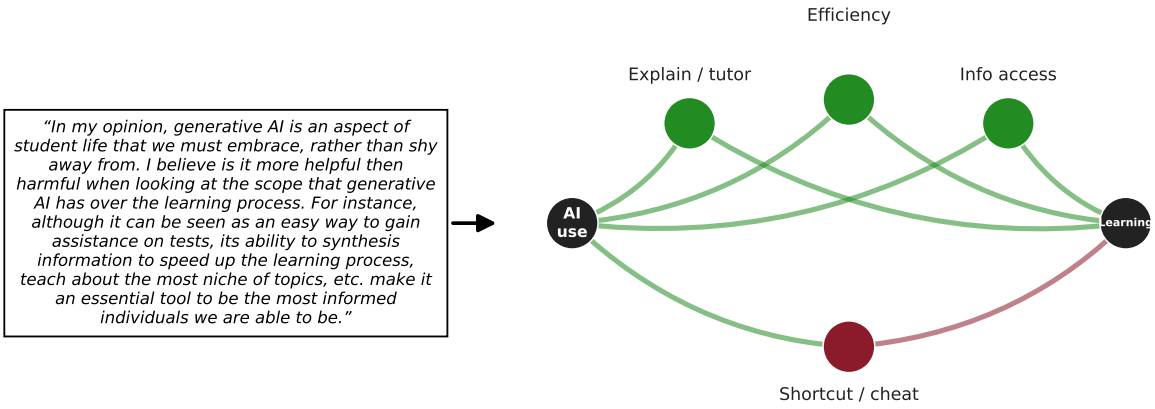
We proceed in two steps. First, we develop the codebook. Using the full set of student responses, we catalog the distinct mechanisms students raise, and consolidate near-synonyms into fifteen mechanisms. Second, we code every response against this fixed codebook with an LLM. We group the mechanisms by their effect on learning. Seven promote it: explanation and tutoring, efficiency, information access, brainstorming, personalized feedback, access and equity, and engagement. Eight harm it: shortcutting and cheating, cognitive offloading, skill atrophy, dependency, inaccuracy, reduced productive struggle, retention loss, and reduced human contact. A final node, *mode of use*, marks responses that make the effect explicitly conditional ("it depends on how you use it"). This node is a mediator between AI use and the channels rather than a channel itself: AI use feeds mode of use, which branches into channels that raise learning and channels that lower it, so conditionality is encoded by the node carrying both a promoting and a harming arm.

To illustrate how the coding works, Appendix Figure C3 shows the directed acyclic graphs of three specific students. The coder assigns each response an overall valence: the student's bottom-line judgment of whether AI helps learning (Panel A), harms it (Panel B), or depends on how it is used (Panel C). The Panel A student concludes that AI helps learning yet flags the temptation to shortcut; the Panel B student concludes that AI harms learning yet credits AI with explaining difficult concepts. The pattern extends beyond these two students: 20 percent of net-positive narratives name at least one harming channel, and 27 percent of net-negative narratives name at least one promoting channel.

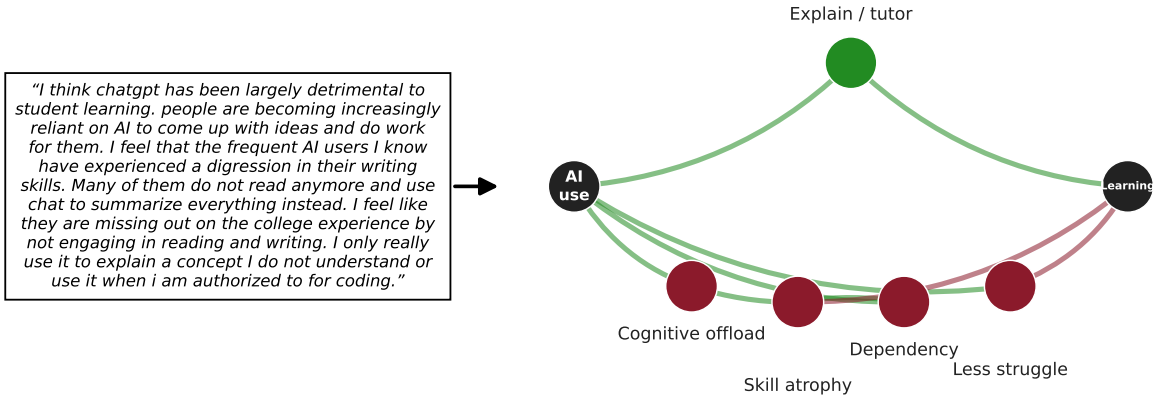
The taxonomy also maps onto the use classification in the main text. Explanation, feedback, and brainstorming are channels in which AI works *with* the student; shortcutting and offloading are channels in which AI works *for* the student. The coder accordingly classifies the manner of use each response describes: *augmentation* when the student describes working with AI (asking it to explain, give feedback, or check work), *automation* when the student describes AI doing the work in their place, both, or unspecified.

Figure C3: Three Example Student Narratives, Coded as Signed Causal Graphs

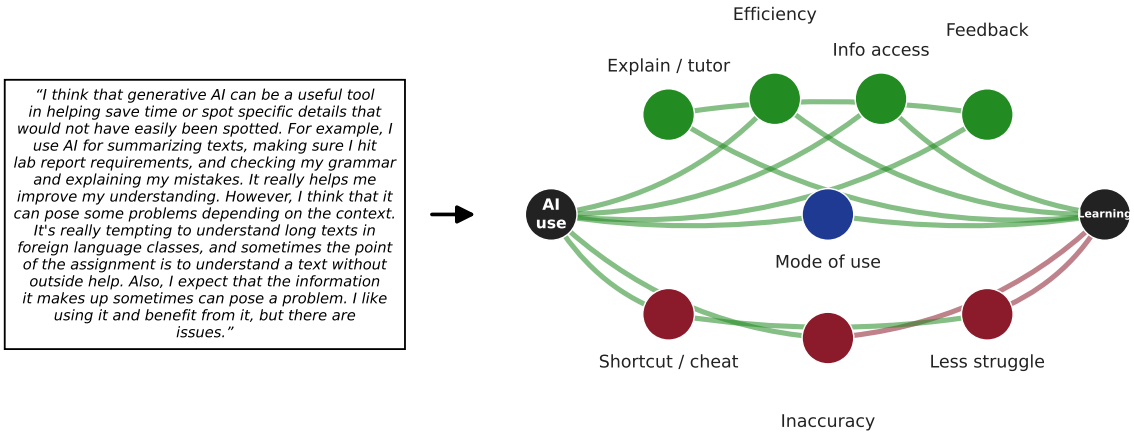
Panel A. A narrative in which AI helps learning



Panel B. A narrative in which AI harms learning



Panel C. A conditional narrative



Notes: Each panel pairs a student's verbatim open-ended response (left) with the response coded as a signed causal graph from AI use to learning (right), following Andre et al. (2026), so the mapping from text to graph can be assessed directly. Green links denote mechanisms the student describes as promoting learning, maroon links mechanisms that harm it, and navy the mode-of-use node.

We code each response with an LLM (Claude Opus 4.8). The majority of responses (98 percent) can be represented as a causal graph; the rest are non-substantive (e.g., "I don't know"). As a validation, we code every response a second time with a model from a different developer (OpenAI's GPT-5.5) and compute their inter-rater agreement statistic, the pooled probability that a code assigned by one rater is also assigned by the other. Agreement matches the trained human coders in Andre et al. (2026): 0.88 for the individual mechanisms (0.88 in their data), 0.97 for the coarser promotes-versus-harms grouping (0.94 and 0.93), and 0.73 for the signed links (0.77), with the mechanism and link statistics far above their random-coding benchmarks of 0.32 and 0.21, respectively. The two models agree on a response's overall valence 94 percent of the time and on its augmentation-versus-automation framing 85 percent of the time.[28]

C.3 The Narratives of College Students

The Average Narrative. The average narrative contains 3.6 mechanisms and 6.4 links, and 96 percent of coded responses mention more than one mechanism. Among students naming any mechanism, 69 percent name both a promoting and a harming channel, and 69 percent of coded responses frame AI's effect as explicitly conditional. The most common single node is mode of use (54 percent), and the most common promoting and harming mechanisms are explanation and tutoring (40 percent) and shortcutting and cheating (42 percent).

Variation Across Students. We ask how narratives vary with student characteristics. Appendix Table C1 regresses each of six narrative features—three *sophistication* measures (the number of mechanisms, the number of signed links, and an indicator for naming more than one mechanism) and the three *framing* measures (augmentation, automation, and conditional-on-use)—on the randomized treatment and the full set of student characteristics from Appendix Table A1.

Among the AI-experience measures, only self-assessed proficiency predicts how students frame AI. Students with above-median proficiency are 17.0 percentage points more likely

[28]Two checks confirm that the coding is not model-specific. Coding every response a third time with a different model from the primary coder's developer (Anthropic's Claude Sonnet 4.6) gives nearly identical agreement with the primary coder (0.86 for mechanisms, 0.66 for links, and 0.92 for valence). Agreement declines with model capability: within the OpenAI line, GPT-5.5 reaches mechanism and link agreement of 0.88 and 0.73, whereas two smaller models reach 0.76 and 0.52, and 0.70 and 0.47. All remain far above the random-coding benchmarks, but we use frontier models for both coders.

than less proficient students to frame AI as augmentation ($p = 0.019$) and 11.9 pp less likely to frame it as automation ($p = 0.065$). Early adopters—those who first used AI for academics before Fall 2024—do not differ from later adopters on any narrative feature, and randomized AI access shifts none of the six: the largest treatment difference, a 7.2 pp lower automation share among treated students, is imprecisely estimated ($p = 0.205$). Framing thus tracks how well students use AI, not how often they use it, how early they began, or whether they received a single session of access.

Demographics, academic background, and beliefs, by contrast, do predict narratives. Men name 0.39 fewer mechanisms and 0.74 fewer links than women (both marginally significant; $p = 0.093$ and $p = 0.073$), the same gender difference Andre et al. (2026) document in narratives about inflation. Students with above-median GPAs are 8.8 pp more likely to name more than one mechanism ($p = 0.037$). Framing varies instead with field of study and beliefs about AI. Natural-science majors are 29 pp more likely to condition AI's effect on how it is used ($p = 0.002$) and 19.9 pp less likely to frame it as automation ($p = 0.004$). Students who expect large AI effects on their own test scores frame AI the same way: they are 13.0 pp less likely to frame it as automation ($p = 0.028$) and 15.1 pp more likely to condition its effect on use, though the latter is marginally significant ($p = 0.063$).

Table C1: Correlates of Students' Narratives

	Narrative sophistication			Use-mode framing		
	# mech.	# links	>1 mech.	Aug-ment.	Au-tomat.	Condit.
	(1)	(2)	(3)	(4)	(5)	(6)
AI access (randomized)	0.083	0.297	−0.044	−0.013	−0.072	0.055
	(0.241)	(0.449)	(0.033)	(0.060)	(0.057)	(0.076)
Panel A. Demographic characteristics						
Male	−0.387*	−0.740*	−0.026	−0.062	0.010	−0.035
	(0.229)	(0.411)	(0.032)	(0.056)	(0.058)	(0.073)
White	0.163	0.422	−0.026	0.087	−0.022	−0.107
	(0.228)	(0.405)	(0.036)	(0.069)	(0.062)	(0.076)
International	−0.204	0.025	0.000	0.081	−0.040	−0.089
	(0.308)	(0.559)	(0.042)	(0.089)	(0.079)	(0.102)
Private high school	0.136	0.085	−0.002	0.013	−0.020	0.011
	(0.243)	(0.442)	(0.037)	(0.066)	(0.071)	(0.084)
Panel B. Academic background						
Natural Sciences	−0.101	0.082	0.025	−0.106	−0.199***	0.291***
	(0.274)	(0.487)	(0.053)	(0.072)	(0.067)	(0.093)
Social Sciences	0.003	0.221	0.042	0.041	−0.065	0.060
	(0.273)	(0.473)	(0.039)	(0.069)	(0.074)	(0.091)
High GPA	0.391	0.643	0.088**	−0.011	−0.006	0.011
	(0.243)	(0.431)	(0.042)	(0.069)	(0.063)	(0.079)
Freshman	0.567**	0.971*	0.003	−0.113	−0.052	0.143
	(0.286)	(0.506)	(0.050)	(0.073)	(0.078)	(0.092)
Sophomore	0.351	0.669	0.041	0.035	−0.058	0.078
	(0.380)	(0.687)	(0.048)	(0.114)	(0.085)	(0.115)
Junior	0.481	0.797	0.051	−0.108	0.020	0.029
	(0.305)	(0.537)	(0.034)	(0.076)	(0.079)	(0.097)
Panel C. AI experience and beliefs						
Frequent AI user	0.319	0.610	−0.034	0.008	−0.078	0.082
	(0.282)	(0.497)	(0.033)	(0.073)	(0.059)	(0.078)
Early AI adopter	−0.198	−0.294	−0.002	0.031	0.003	0.005
	(0.249)	(0.434)	(0.033)	(0.059)	(0.055)	(0.068)
High AI proficiency	0.115	0.313	−0.036	0.170**	−0.119*	−0.028
	(0.291)	(0.522)	(0.029)	(0.072)	(0.064)	(0.084)
High perceived AI effect	0.462*	0.765	0.016	−0.048	−0.130**	0.151*
	(0.258)	(0.480)	(0.042)	(0.067)	(0.059)	(0.081)
N (Students)	200	200	200	200	200	200

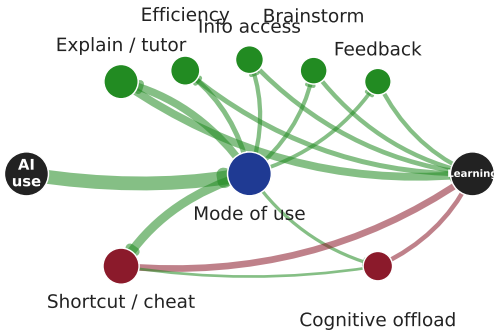
Notes: Each column reports an OLS regression of the indicated narrative feature on all rows jointly, estimated on the codable open-ended responses. Outcomes come from each student's coded causal graph: the number of mechanisms, the number of signed links, an indicator for naming more than one mechanism, and indicators for the augmentation, automation, and conditional-on-use framings. All characteristics are indicators, defined as in Appendix Table A1; missing values are set to zero, with missing indicators included. Heteroskedasticity-robust standard errors in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

Narrative Clusters. We next ask whether students' narratives come in recurring types or "clusters." Following Andre et al. (2026), we measure the distance between two narratives as the fraction of their combined signed links that the two do not have in common (the Jaccard distance), group narratives with agglomerative hierarchical clustering with average linkage, and choose the number of clusters to maximize the average silhouette, a standard measure of how well each narrative fits its own cluster relative to the nearest alternative.

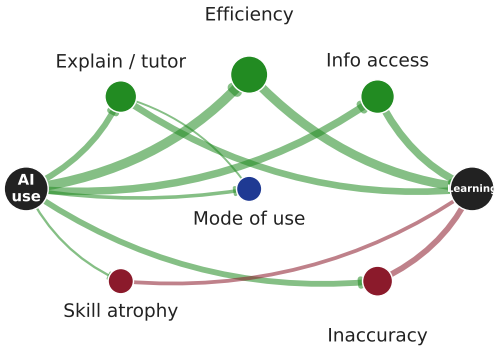
Four narrative types cover 94.0 percent of coded responses, while the rest scatter across small clusters (Appendix Figure C4). The modal type (45 percent) is the double-edged narrative, in which mode of use feeds both an explanation channel and a shortcutting channel. An augmentation type (32 percent) chains efficiency, information access, and explanation to higher learning, though many of these narratives also flag accuracy concerns, and an automation type (10 percent) chains shortcutting and offloading to lower learning. The remaining type is smaller and one-sided: a deskilling narrative (7.0 percent), in which skill atrophy and offloading erode critical thinking.

Figure C4: Narrative Clusters

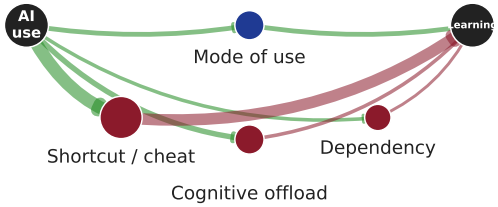
Panel A. Double-edged (45 percent)



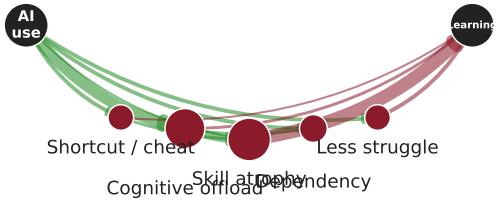
Panel B. Augmentation (32 percent)



Panel C. Automation (10 percent)



Panel D. Deskilling (7.0 percent)



Notes: This figure shows the average causal graph within each narrative cluster; panel headings report each cluster's share of coded narratives. We form the clusters with agglomerative hierarchical clustering, using average linkage on the Jaccard distance between signed-link sets and choosing the number of clusters to maximize the average silhouette. We show the four clusters with at least eight responses and, within each panel, omit mechanisms below 20 percent and links below 6 percent.