Experimental Evidence on the Learning Impact of Generative AI*

Zara Contractor Germán Reyes

March 2026

Abstract

We estimate the effects of generative AI on learning through a randomized experiment with 210 undergraduate students. Participants learn a new topic with or without AI access, and we measure outcomes both immediately and one week later. AI access improves test scores by 0.25 standard deviations, with gains concentrated among middle-performing students and persisting one week later. AI access also changed how students write: treated students produced longer essays with shorter sentences and simpler vocabulary that independent graders rated as higher-quality. We find that the learning benefits were driven by a reallocation of time from writing toward information-seeking activities and by greater enjoyment of the learning task.

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

1 Introduction

Generative artificial intelligence (AI) tools like ChatGPT are changing how college students acquire human capital. These tools could enhance learning by acting as an on-demand tutor that explains difficult concepts and provides personalized feedback, but also undermine it by substituting for the cognitive effort that learning requires—writing entire essays, completing problem sets, and generating code with minimal student input. This tension has sparked debate about how AI is affecting student learning, with positions relying largely on anecdotal evidence and speculation (Herrman, 2025; Hsu, 2025).

In this paper, we estimate the immediate and longer-term effects of AI access on learning through a randomized experiment with 210 undergraduate students. The experiment consists of two in-person sessions conducted approximately one week apart. In the first session, participants are randomly assigned to either an AI-allowed or an AI-forbidden condition and asked to learn about a randomly assigned topic and write an essay about it; the AI-allowed group may use AI tools during this learning phase. All participants then complete a knowledge test without access to AI or other external resources. In the second session, approximately one week later, students complete another essay and test, both without AI access.

We first document that our random assignment generated substantial variation in AI usage. Nearly 70 percent of students in the AI-allowed condition used generative AI during the learning phase, compared to near-zero usage in the control group. Students primarily used AI for conceptual support rather than text generation: 44 percent used it to explain difficult concepts, 23 percent to summarize readings, and only 16 percent to draft responses. Among AI users, 88 percent reported that the tools were helpful for the learning task.

Our main finding is that AI access improved learning outcomes. Students with AI access scored 0.25 standard deviations higher on the immediate knowledge test, a 6.4 percentage point increase in the share of questions answered correctly. This improvement was concentrated among students in the middle of the performance distribution, with smaller effects at the extremes. The gains persisted approximately one week later, though with some attenuation: students who had used AI during initial learning scored 0.20 standard deviations higher on the retention test, about 81 percent of the initial effect.

We also find that AI access changed how students write. Treated students produced essays that were about seven percent longer, with shorter sentences and simpler vocabulary, and these essays were rated higher-quality by independent graders. Commercial AI content

detectors flag a significantly higher fraction of treated students' text as AI-generated, though AI-allowed essays were no more "homogeneous" than control essays, contrasting with the textual convergence documented in workplace settings (Brynjolfsson et al., 2025). These writing changes partially persisted one week later, when students wrote without AI access, suggesting that exposure to AI-generated text may have lasting effects on students' own writing.

To examine mechanisms, we analyze how AI access affected student behavior during the learning phase. AI access did not change total time spent on the task or engagement with provided materials, but it reallocated time across activities: students spent less time writing and more time reading and searching for information. AI access also substantially increased enjoyment of the learning task, with treated students reporting 11 percent higher enjoyment ratings. Finally, AI access during the learning phase increased academic integrity violations during the test, suggesting that exposure to AI may erode compliance with rules prohibiting its use.

This paper contributes to three literatures. First, we contribute to the literature on productivity effects of generative AI. Field experiments document substantial productivity gains from AI access in professional writing, software development, and customer support (Noy and Zhang, 2023; Peng et al., 2023; Dell'Acqua et al., 2023; Cui et al., 2024; Brynjolfsson et al., 2025). Yet most productivity studies measure *skill deployment*—how well workers perform tasks using existing knowledge—rather than *skill accumulation*—how effectively individuals acquire new knowledge. The few studies examining learning within work settings find mixed results: Brynjolfsson et al. (2025) show that AI access accelerates skill acquisition among call-center workers, while Budzyń et al. (2025) find no such gains among endoscopists. We provide experimental evidence that AI access can improve skill accumulation in an academic context, identifying a channel through which AI may raise long-run productivity beyond its immediate effects on task performance.

Second, we advance the emerging literature on generative AI in education. A large body of work documents how students adopt and perceive AI tools (e.g., Stöhr et al., 2024; Ravšelj et al., 2025), but adoption patterns reveal little about learning impacts. Causal evidence remains scarce because exogenous variation in AI access is hard to generate, and existing experiments point to nuanced effects that depend on implementation details. Bastani et al. (2025) find that unrestricted ChatGPT access harms learning in high school mathematics, though guardrails preventing direct solution provision mitigate these effects. Lehmann et al. (2024) show that in coding courses, students who use LLMs to seek ex-

planations learn more, while those who copy-paste solutions learn less. We provide causal estimates of unrestricted AI access on both immediate learning and knowledge retention one week later, and unpack the behavioral mechanisms—time allocation, engagement, and academic integrity—through which these effects operate.

Third, we contribute to the broader literature on digital technologies and student learning. Prior research has explored laptops and personal computers (Carter et al., 2017; Malamud and Pop-Eleches, 2011; Cristia et al., 2017), online courses (Figlio et al., 2013; Bettinger et al., 2017), internet access (Vigdor et al., 2014; Dettling et al., 2018), and interactive technologies such as clickers and whiteboards (Caldwell, 2007; Lewin et al., 2008). This literature finds mixed results: generic technology provision often yields limited or negative effects, while interactive and personalized technologies show more consistent positive impacts (see Bulman and Fairlie, 2016; Escueta et al., 2020, for reviews). Generative AI is a fundamentally different educational technology—one that can engage in natural language dialogue, provide personalized explanations, and generate novel content tailored to individual queries. Our experimental design isolates the causal effect of this technology on learning outcomes while examining the behavioral mechanisms through which these effects operate.

2 Context: Use of Generative AI in Higher Education

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

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

Ex ante, predicting how AI will affect learning outcomes is difficult because generative AI is a general-purpose technology that students employ for many purposes. One useful framework groups these uses into augmentation versus automation (Autor and Thompson, 2025). Augmentation includes obtaining explanations of difficult concepts or receiving feedback on drafts. Automation includes generating essays or problem set solutions. If students primarily use AI to automate cognitive tasks, it may reduce the effortful processing necessary for durable learning. Conversely, if students primarily use AI to augment their understanding, it may enhance learning outcomes. Contractor and Reyes (2025) find that students use AI for both purposes, a pattern consistent actual usage data from Claude conversation logs (Handa et al., 2025) and from an analysis of ChatGPT interaction logs among undergraduate students at another U.S. university (Ammari et al., 2025), making the net impact on learning theoretically ambiguous.

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

3 Experimental Design

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

3.1 Session One: Knowledge Production and Initial Learning

Overview. The first session employs a between-subjects design with random assignment to either an AI-allowed or AI-forbidden condition. We scheduled eight time slots with two concurrent sessions per slot, one for each treatment arm, conducted in separate computer labs. For each time slot, we randomly assigned one lab to the AI-allowed condition and the other to the AI-forbidden condition. Students who signed up for a given time slot were randomly assigned to a specific computer in one of these two labs. All computer workstations were equipped with privacy dividers to minimize distractions and prevent participants from viewing other screens (see Appendix Figure A1).[1]

[1]We note four main deviations from the experimental protocol that occurred during early sessions. First, the initial Session One verbal welcome script did not mention AI permissions or restrictions; participants learned their treatment condition only through written platform instructions. We introduced verbal acknowledgment of treatment conditions starting with the second Session One round. Second, the initial Session One included a timer that malfunctioned, displaying available time after the 35-minute limit had

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

Participants in the *AI-allowed* condition receive explicit instructions permitting use of any generative AI tools during the learning phase. These instructions are embedded in the welcome script that proctors read aloud at the start of each session. The script states: "[I]f you typically use generative AI tools such as ChatGPT, feel free to use them here as well. Using ChatGPT or other generative AI is completely allowed. We provide you with a ChatGPT account, already logged in one of the tabs, so you don't need to use your personal account." The instructions are also communicated in written form through the survey interface (Appendix Figure A2, Panel A). To facilitate usage, each computer has three browser tabs already open: a dedicated ChatGPT (GPT-4o) account, Wikipedia, and the Middlebury College Library website.

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

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

elapsed. We removed the timer for subsequent sessions. Third, the first Session Two round lacked a waiting room, requiring participants to manually track time before leaving. We introduced a virtual waiting room for all following sessions. Fourth, a technical glitch in one Session Two prevented participants from completing the essay on the platform. We added extra time and asked affected participants to write their essays in Microsoft Word, manually incorporating these responses into the dataset. All analyses include session fixed effects to account for these protocol variations.

from external sources (Noy and Zhang, 2023).

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

Welcome and Instructions. Each session begins with a staff member reading a condition-specific welcome script. We tell participants that their primary task involves learning about a pre-specified topic and that they will demonstrate their understanding through a series of assessments. To ensure comprehension of the experimental procedures, participants must correctly answer a series of understanding checks, with laboratory staff providing additional clarification if needed. Participants answered an average of 4.76 out of 5 comprehension questions correctly, and the median participant answered all questions correctly.

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

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

Learning Phase. Following the baseline test, participants enter the main learning phase. During this phase, they have up to 35 minutes to learn about their assigned topic and write an approximately 500-word analytical essay demonstrating their understanding. 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).

The writing assessment consists of an analytical task examining specific aspects of the assigned topic. Participants are asked to analyze relationships between concepts, compare different elements, and support their analysis with specific examples (Appendix B.2 contains complete prompt texts). For each topic, we developed two different prompts and randomly assigned each participant one prompt for Session One and the other prompt in Session Two.

We provide all participants with a link to an introductory reading about their assigned topic. We constructed these materials to be of comparable length and reading difficulty across all three topics. Texts range from 1,253 to 1,399 words in length (Appendix Ta-

ble B1). Based on an average silent reading speed of approximately 250 words per minute for college students (Carver, 1992; Brysbaert, 2019), the estimated reading time for each text is 5.0 to 5.6 minutes. We track whether participants click on this link as a measure of engagement with provided resources. The full text of the learning materials is available in the Supplementary Materials.

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

Writing assessments are evaluated using both human graders and AI grading. Human graders are independent evaluators recruited through Prolific who are blind to treatment condition and participant identity. Each essay is independently evaluated by three to four human graders, who assess the overall quality of the written work on a 0-10 scale, with 10 representing excellent understanding and synthesis of the topic (see Appendix B.3 for the grading instrument). In addition to overall quality, other evaluation dimensions include comprehension (accurate understanding of key concepts), analytical reasoning (depth of analysis and critical thinking), organization (logical flow and structure), and evidence use (effective integration of factual information). We complement human grading with AI grading using Claude Opus, which evaluates each essay on the same six dimensions using the same rubric.[2] In addition to these quality ratings, we conduct a linguistic analysis of the text, measuring length, lexical diversity, and readability. This allows us to assess how AI access influences written expression beyond subjective quality ratings.

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

Post-Learning Test. The session concludes with five multiple-choice questions testing factual knowledge and conceptual understanding of the assigned topic. These questions differ from the baseline test. Participants complete this assessment without access to any

[2]Our use of AI grading is motivated by a growing literature validating LLMs as evaluators of text quality (Zheng et al., 2023; Chiang and Lee, 2023; Liu et al., 2023), and by the fact that our essay topics (blockchain, carbon capture, CRISPR) are technical subjects on which human graders recruited from a general population may have limited expertise. The correlation between average human and AI scores for overall essay quality is 0.48 in Session One and 0.57 in Session Two (both $p < 0.001$). See Appendix B.4 for the AI grading procedure and prompts.

external resources, internet, or AI tools regardless of their treatment condition.

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

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

3.2 Session Two: Knowledge Retention

Session Two takes place approximately one week after Session One.[3] This session measures knowledge retention when all participants—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).

Knowledge Retention Test. One of the two assessments is a 10-item multiple-choice test covering factual content from the randomly-assigned topic. We designed these questions to be distinct from those in Session One.

Analytical Essay. The second assessment is a 20-minute analytical writing task on the assigned topic. This task uses the complementary prompt we developed for each topic. Independent graders evaluate these submissions using the same rating criteria and procedures established in Session One.

Exit Survey. Session Two concludes with a questionnaire measuring beliefs about AI's impact on learning, understanding, course grades, and time spent on academics; self-

[3]When scheduling, we encouraged participants to sign up for a Session Two slot approximately one week after their Session One slot, but allowed flexibility to accommodate scheduling constraints. The average gap between sessions was 6.9 days. Over 60 percent of students scheduled their second session exactly 7 days after Session One, and over 80 percent attended within 6 to 8 days.

reported AI skills (Likert scale) and frequency of AI use for different academic tasks (never to daily); when participants started using AI and specific academic applications (brainstorming, writing, editing, research, coding, etc.); retrospective Session One behavior including self-reported use of AI during Session One, which tools were used, and for what purposes (explaining concepts, summarizing readings, drafting, proofreading, editing); whether participants studied the topic or accessed materials between sessions; and an open-ended question about the impact of generative AI on learning.

4 Data and Research Design

4.1 Sample and Recruitment

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

Students were required to attend two in-person sessions approximately one week apart. We implemented several design features to minimize attrition. First, we structured compensation to incentivize completion of both sessions: participants receive $5 for completing Session One alone, but $50 for completing both sessions, with all payments disbursed at the conclusion of Session Two. Second, participants received email and text reminders on the day of their scheduled session. Third, we provided flexibility in the timing of the second session rather than requiring exactly seven days between sessions. We also allowed students to reschedule if conflicts arose.

In total, 256 students completed the intake survey and reserved a Session One slot. Of these, 211 students (82.4 percent) attended Session One, and 204 students (79.7 percent) attended both sessions. We drop one Session One participant due to incomplete data, yielding an analysis sample of 210 students. Among students who attended Session One, 97.1 percent returned for Session Two.

4.2 Summary Statistics and Balance

Table 1 reports summary statistics. Column 1 includes students who completed the intake survey and signed up for a Session One slot. Columns 2 and 3 include students who at-

tended Session One and both sessions, respectively. We focus on students who attended Session One (column 2). The typical student is 20.1 years old; 35.2 percent are male, 49.5 percent white, 26.7 percent international, and just over half (53.8 percent) attended a public high school (Panel A). The sample spans all four undergraduate years, with 37.6 percent first-years, and represents diverse academic backgrounds: 30.5 percent study natural sciences, 35.7 percent social sciences (Panel B). Academic performance is strong, with a mean GPA of 3.68 and an average SAT score of 1386, approximately the 93rd percentile of test takers (College Board, 2025).[4] Students self-report little baseline knowledge of their assigned topic (1.5 on a 0–10 scale) and, consistent with this, answer only 30.4 percent of baseline multiple-choice questions correctly.

The sample is balanced on most observable characteristics across treatment groups. Table 2 compares mean characteristics of students in the AI-forbidden versus AI-allowed conditions (columns 1 and 2). While some differences are statistically significant—notably age and the likelihood of having already declared a major—they are small in magnitude (column 3). An F-test of joint balance cannot reject the null that all differences equal zero at conventional levels. Most importantly, measures of academic ability—strong predictors of learning outcomes—are well-balanced: mean GPA (3.69 versus 3.66) and SAT scores (1392 versus 1382) are nearly identical across groups. To address any residual imbalance, we use the double-lasso procedure (Belloni et al., 2014), as described below. We find no differential attendance in Session One (82.5 versus 82.2 percent) and no differential attrition between sessions.

4.3 Regression Model

We estimate linear models of the form:

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

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

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

outcome or the treatment assignment from a pool of potential covariates.[5] We report heteroskedasticity-robust standard errors.

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

4.4 Main Outcomes

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

AI Usage. As a measure of a "first-stage", we analyze whether students used generative AI during the learning phase through multiple complementary indicators. The revealed preference measure equals one if at least one prompt was found in the ChatGPT account provided to the student during Session One or if a proctor directly observed the student using ChatGPT. Self-reported measures capture whether students used any generative AI tool, whether they specifically used ChatGPT, and whether they used alternative AI models (Claude, Gemini, Copilot, DeepSeek, or Llama).

Test-Based Learning Outcomes. We measure knowledge acquisition and retention using both self-reported and objective measures, as each has complementary strengths. Self-assessed knowledge captures students' subjective evaluation of their understanding on a 0-10 scale, where 0 indicates "I know nothing about this topic" and 10 indicates "I am an expert." This measure may capture dimensions of learning that a limited set of multiple-

[5]The pool of potential controls includes demographic characteristics (age, gender, race/ethnicity indicators for Black, Latino, Asian, international student status, public versus private high school attendance), academic background (cohort fixed effects, declared major status, field of study indicators for natural sciences, social sciences, and humanities/arts, self-reported study hours per week, college GPA, SAT score, whether an admissions test was taken, plus binned versions of SAT scores, college GPA, and high school GPA to allow for flexible functional forms), baseline measures (pre-learning self-assessed knowledge, pre-learning test score, prior experience with similar tasks, comprehension check score), and experimental design features (topic fixed effects, writing prompt version, and lab location).

choice questions cannot test, but it requires accurate self-knowledge and is subject to the usual biases of subjective questions, such as social desirability bias and reference bias. The fraction correct measure represents the proportion of multiple-choice questions answered correctly (five questions in Session One, ten questions in Session Two). This objective measure overcomes the biases inherent in self-reports but is constrained by the number of questions, which limits our ability to test learning across all relevant dimensions. We construct standardized test scores by normalizing raw scores to have mean zero and standard deviation (SD) one in the control group. We additionally define binary indicators for scoring above various performance thresholds, which allow us to examine treatment effects across the score distribution.

Writing Quality. While multiple-choice tests measure factual recall, writing assessments capture higher-order skills such as analytical reasoning, synthesis, and argumentation. We view the grader-assessed essay scores described above as our primary measure of critical thinking. Each essay was independently evaluated by three to four graders, and we use the individual grader-essay pair as the unit of observation, including grader and batch fixed effects to absorb differences in grader leniency. We analyze each of the six dimensions: overall quality, factual accuracy, use of evidence and examples, relevance to the prompt, structure and organization, and writing style and clarity. We also construct binary indicators for scoring above various quality thresholds to examine distributional effects.

Writing Style. We analyze multiple dimensions of participants' written output. Essay length is measured as the total number of tokens. We also create an indicator for empty essays, which equals one if the student submitted no written response. We construct several readability measures: the Flesch Reading Ease Score (0-100 scale, higher values indicate easier text), the Flesch-Kincaid Grade Level (years of formal education required to comprehend the text), and mean sentence length (average words per sentence). We measure lexical diversity through the Type-Token Ratio (ratio of unique words to total words) and the proportion of hapax legomena (share of words appearing only once).

We also examine whether AI access homogenizes student writing, following the textual convergence analysis in Brynjolfsson et al. (2025). We compute sentence embeddings for each essay using the `all-mpnet-base-v2` transformer model and measure within-group textual similarity as the average pairwise cosine similarity between a student's essay and all other essays written by students in the same treatment group, topic, and essay prompt.[6]

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

5 Immediate Effects of AI Access on Learning and Writing

5.1 First Stage: AI Adoption and Usage Patterns

We begin by examining AI access's impact on AI usage during the learning phase. Table 3 reports estimates from equation (1) using several measures of AI usage. The first row presents the revealed preference measure based on activity in the ChatGPT accounts provided to treatment group participants combined with direct proctor observation. The remaining rows present self-reported measures collected in Session Two: use of any generative AI tool, use of ChatGPT specifically, and use of other AI models (Claude, Gemini, Copilot, and others). 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 66.3 percentage points (pp) as measured by activity in the provided ChatGPT account ($p < 0.01$). This represents a large increase from the near-zero baseline in the control group. Self-reported measures yield similar though slightly attenuated results: overall generative AI usage increased by 59.9 pp ($p < 0.01$). Virtually all adoption occurred through ChatGPT—usage increased by 60.2 pp ($p < 0.01$), while usage of other AI models remained negligible at 0.9 pp.[7]

Having established that AI access substantially increased usage, we examine how treated students used these tools. Figure 3 presents the frequency of different AI applications among AI users. Students thus predominantly used AI for conceptual support—understanding and synthesizing information. The most common use was explaining concepts (44.3 percent), followed by summarizing readings (22.6 percent) and drafting responses (16.0 percent). Notably, these tasks do not require AI—e.g., students can find online explanations of technical concepts—yet having access to AI proved valuable. Over 88 percent of students in the AI-allowed group reported that AI was somewhat or very helpful for the learning task.

We examine which student characteristics predict AI take-up among those assigned to the treatment group (Appendix Table A1). Students who report more frequent AI use during the academic semester and higher self-assessed AI proficiency were significantly more likely to use the provided ChatGPT account (28.4 pp and 18.3 pp, respectively; both $p < 0.05$). These associations persist in multivariate regressions that control for all characteristics simultaneously. Students who perceive larger benefits from AI—measured

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

as the difference between their believed performance with and without AI access—are significantly more likely to self-report AI use (23.4 pp, $p < 0.05$), though this association does not appear in the revealed-preference measure. Early AI adoption, demographic characteristics, and field of study do not significantly predict take-up. The strong role of prior AI experience suggests that familiarity with AI tools lowers barriers to adoption in new contexts.

5.2 Effects on Writing Style

Next we examine how AI access affected participants' writing style during Session One. Table 5 presents ITT and TOT estimates for five writing indices: length, readability, lexical diversity, within-group textual similarity, and source text similarity. Figure 5 presents visual evidence on these treatment effects in standard deviation units.

AI access suggestively led students to produce longer essays written in a clearer, more accessible style—using simpler words and shorter sentences. The length index increased by 0.15 SD (column 2), corresponding to about 15 additional tokens, or approximately a seven percent increase from the control mean of 203 tokens (Appendix Table A2). The readability index increased by 0.09 SD (column 2), reflecting improvements across multiple dimensions: the Flesch Reading Ease score increased by 3.8 points and mean sentence length fell by 4.6 words (Appendix Table A2). The lexical diversity index shows a small positive effect of 0.08 SD (column 2). While these point estimates are imprecisely estimated and do not reach statistical significance, they suggest a pattern of AI-facilitated changes in writing style toward greater clarity and accessibility.

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

Finally, we examine whether AI access left detectable traces in students' writing using Pangram, a commercial AI content detector. The fraction of text classified as AI-generated increased by 10.6 pp in the AI-allowed group ($p < 0.05$), an 83 percent increase over the control mean of 12.8 percent. Notably, the control mean itself is non-trivial: 12.8 percent of control group text is flagged as AI-generated despite these students having no access to

AI tools. This baseline reflects either false positives or students whose natural writing style resembles AI-generated text—a pattern consistent with recent evidence that exposure to large language models has begun shifting human communication patterns toward AI-like prose (Yakura et al., 2024). In Session Two, when no students had AI access, the treatment effect drops to zero, and the control mean falls to 4.1 percent, suggesting that the Session One flags in the control group partly reflected topic-specific or prompt-specific stylistic patterns rather than stable individual writing traits.

5.3 Effects on Writing Quality

We next examine whether AI access affected the substantive quality of students' essays, as assessed by independent graders blind to treatment condition. Table 4 reports treatment effects on each of the six grading dimensions, standardized to have mean zero and standard deviation one in the control group. Figure 4, Panel A presents ITT effects visually in standard deviation units.

AI access had no measurable impact on essay quality across all six dimensions, with the point estimates being small in magnitude and not statistically significant. The largest improvements appear in the presentation dimensions: writing style and clarity, and structure and organization. Substantive dimensions—factual accuracy, use of evidence, and relevance to the prompt—also show positive point estimates. Overall quality, the grader's holistic assessment, shows a similar pattern. Appendix Table A3 examines distributional effects on overall essay quality. In Session Two, AI access increased the probability of scoring 9 or above by 5.5 pp ($p < 0.05$), suggesting that AI exposure may have helped students at the margin of producing high-quality work.

5.4 Effects on Learning

Next, we examine the impact of AI access on immediate learning outcomes measured at the end of Session One. AI access produced significant improvements in actual learning but not in perceived learning (Table 6). Self-assessed knowledge decreased by 0.02 points (column 1), an effect that is not statistically significant. Performance on the objective knowledge test improved substantially. Students in the AI-allowed condition answered 6.4 pp more questions correctly (column 2, $p < 0.05$), equivalent to an 11 percent improvement over the control mean of 56.3 percent correct. When standardized, this translates to a 0.25 SD improvement in test scores (column 3, $p < 0.05$). The disconnect between

subjective and objective measures suggests that students may not fully perceive the learning gains facilitated by AI access. These effects are similar across all three topics (Appendix Table B2).

The learning gains from AI access are concentrated among middle-performing students. Table 7 reports treatment effects on the probability of scoring above various performance thresholds. Students with AI access were 7.9 pp more likely to score at least 40 percent correct (column 2) and 11.8 pp more likely to score at least 60 percent correct (column 3, $p < 0.10$). AI had minimal effects on reaching the 80 percent threshold (column 4) or achieving perfect scores (column 5). These effects are illustrated in Figure 6, which shows that AI access shifts the middle of the performance distribution rightward while leaving the tails largely unchanged. This pattern is consistent with previous work showing that AI disproportionately benefits workers in the lower half of the performance distribution (Noy and Zhang, 2023; Brynjolfsson et al., 2025).

5.5 Mechanisms: How AI Transforms the Learning Process

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

Engagement. Table 8 examines whether AI access altered engagement with the learning task. Column 1 measures engagement through an indicator for whether students opened the provided reading materials. Columns 2-3 capture time spent on the learning task using both objective data from Qualtrics (recorded time) and participants' self-reported estimates of minutes spent. Column 4 measures topic recall through an indicator for whether participants correctly remembered which of the three topics (blockchain, carbon capture, or CRISPR) they were assigned in Session One.

AI access had no detectable effects on engagement metrics. Access to the reading material did not differ significantly between groups (column 1). Time measures showed no meaningful differences: objectively recorded time decreased by 0.2 minutes (column 2), while self-reported time decreased by 0.6 minutes (column 3), though neither effect is statistically significant. Topic recall was nearly universal and identical across groups: 96 percent of control group students correctly remembered their assigned topic, with a zero treatment effect (column 4).

Time Allocation across Learning Activities. While total learning time remained unchanged, students may have reallocated their time across different activities. Figure 7 plots the treatment effects on time spent across eight learning activities: searching for information online, reading source materials, taking notes, planning the essay, drafting the essay, editing, and browsing unrelated content.

AI access appears to have reallocated time across learning activities, with a suggestive pattern of substitution away from writing and toward information-seeking. Time spent drafting decreased by 1.06 minutes, representing a 10 percent reduction from the control mean of 11.07 minutes. These reductions in writing activities were partially offset by increases in information-seeking: reading time increased by 0.71 minutes, while time spent searching increased by 0.38 minutes, though these effects are imprecisely estimated and not statistically significant at the conventional levels. Other activities showed minimal changes.

Learning Experience. Table 8 examines whether AI access affected participants' subjective experience of the learning task. Columns 1-2 assess task enjoyment on a 0-10 continuous scale and as an indicator for above-median enjoyment. Columns 3-4 similarly measure perceived skillfulness or effectiveness using both continuous and binary measures.

AI access substantially increased participants' enjoyment of the learning task. Task enjoyment increased by 0.60 points on the 10-point scale (column 1, $p < 0.10$), representing an 11 percent improvement from the control mean of 5.22. This effect translates to a 14.8 pp increase in the likelihood of reporting above-median enjoyment (column 2, $p < 0.05$). In contrast, AI access had minimal effects on how skilled or effective students felt during the task. Perceived effectiveness increased by only 0.04 points on the continuous scale (column 3) and 4.5 pp on the binary measure (column 4), with neither effect approaching statistical significance. These findings align with Noy and Zhang (2023), who find that ChatGPT access increases task satisfaction by 0.40 SD and self-efficacy by 0.20 SD among professional writers, though this last effect is not statistically significant. We find a similar-sized effect on task enjoyment (0.28 SD) and also a positive but not significant effect on self-efficacy (0.05 SD).

Academic Integrity. Table 8 investigates how AI access influenced academic integrity during the knowledge tests, where all participants were prohibited from using external resources. We examine four measures of unauthorized resource use: whether proctors

caught students cheating (column 1), self-reported unauthorized resource use (column 2), an indicator for either measure detecting cheating (column 3), and an indicator for both measures detecting cheating (column 4).

AI access during the learning phase significantly increased the likelihood of rule violations during subsequent assessments. Students in the AI-allowed condition were 6.4 pp more likely to be caught by proctors accessing unauthorized resources during tests (column 1, $p < 0.05$), more than doubling the control group rate of 2.9 percent. Self-reported violations increased by 8.6 pp (column 2, $p < 0.05$), more than quadrupling the control mean of 2.0 percent. When combining both detection methods, treatment group students were 12.6 pp more likely to cheat (column 3, $p < 0.01$), representing a large increase from the 5.0 percent baseline.[8]

6 Longer-Term Impacts of Access to Generative AI

6.1 Persistence of Writing Style Changes

We analyze whether the writing style changes observed in Session One persisted when students wrote without AI access one week later. Table 5 and Figure 5 examine the same writing characteristics, now measured in Session Two essays.

Several writing changes persisted even without access to AI, though these effects are imprecisely estimated. The length index increased by 0.19 SD (column 5), corresponding to about 12.8 additional tokens (Appendix Table A2), though this effect is not statistically significant. The readability index showed signs of persistence at 0.05 SD (column 5): essays exhibited easier reading scores (3.7 point increase in Flesch Reading Ease) and shorter sentences (1.8 fewer words per sentence). Lexical diversity decreased by 0.19 SD (column 5, $p < 0.10$): the Type-Token Ratio fell by 0.014 points and the hapax proportion decreased by 0.018 points (Appendix Table A2). While most of these point estimates suggest some persistence in writing style changes, they are not statistically significant at conventional levels, and thus these patterns should be interpreted with caution.

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

6.2 Longer-Term Writing Quality Changes

We examine whether the essay quality improvements from AI exposure persisted when students wrote without AI access in Session Two. Table 4 (columns 4–6) reports treatment effects on the same six grading dimensions, and Figure 4, Panel B presents ITT effects visually.

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

Distributional effects reinforce this interpretation. AI access increased the probability of scoring 9 or above on overall quality by 5.5 pp ($p < 0.05$), relative to a control mean of 7.5 percent (Appendix Table A3). This suggests that AI exposure helped a subset of students produce high-quality essays on their own, consistent with a learning channel rather than mere AI-assisted production.

6.3 Impact on Knowledge Retention

We now examine the effects of AI access during initial learning on knowledge retention measured approximately one week later in Session Two.[9] Table 6 reports estimates using the same outcomes as Session One.

The knowledge advantage from AI access persisted one week later, though with some attenuation. Self-assessed knowledge increased by 0.19 points (column 4), an effect that remains statistically insignificant. Objective test performance remained higher for students who had AI access during initial learning. They answered 3.7 pp more questions correctly (column 5, $p < 0.10$), representing a 7.5 percent improvement over the control mean

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

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

We examine heterogeneity across the performance distribution in Table 7, which reports treatment effects on reaching various performance thresholds. AI access primarily helped students achieve low and medium performance levels. Treatment group students were 4.0 pp more likely to score at least 20 percent correct (column 5). The point estimate suggests a 7.7 pp increase in scoring at least 60 percent correct (column 5), though this effect is not statistically significant. Effects on reaching 40 percent correct, 80 percent correct, or achieving perfect scores are small and statistically insignificant. Figure 6 illustrates these patterns, showing that AI access continues to shift the lower portion of the distribution rightward one week later, though the effects are attenuated compared to Session One.

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

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

After completing the Session Two retention test, participants estimate: (1) how many questions they answered correctly; (2) how many they would have answered under the opposite treatment condition; (3) how many questions other students in their group answered correctly; and (4) the same for students in the opposite group. Figure 8, Panel A compares these beliefs with observed treatment effects; Panel B plots perceived against actual effects across subgroups.

Panel A reveals two facts about individual beliefs. First, students substantially overestimate their own performance: the average student answered 4.9 questions correctly, but control students believed they answered 6.3 and treated students believed 6.7. Second, the degree of overestimation depends sharply on experience with AI. AI-forbidden students predicted that AI would increase their own performance by 2.5 questions—more than six times the actual effect of 0.4 questions—and predicted a 1.4 question increase for others,

more than three times the actual effect. Students who used AI developed far more calibrated beliefs: they predicted a 0.3 question benefit for themselves, closely matching the observed 0.4 question effect. Their predictions about others (1.1 questions) remain inflated but are substantially more accurate than AI-forbidden students' beliefs.

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

7 Conclusion

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

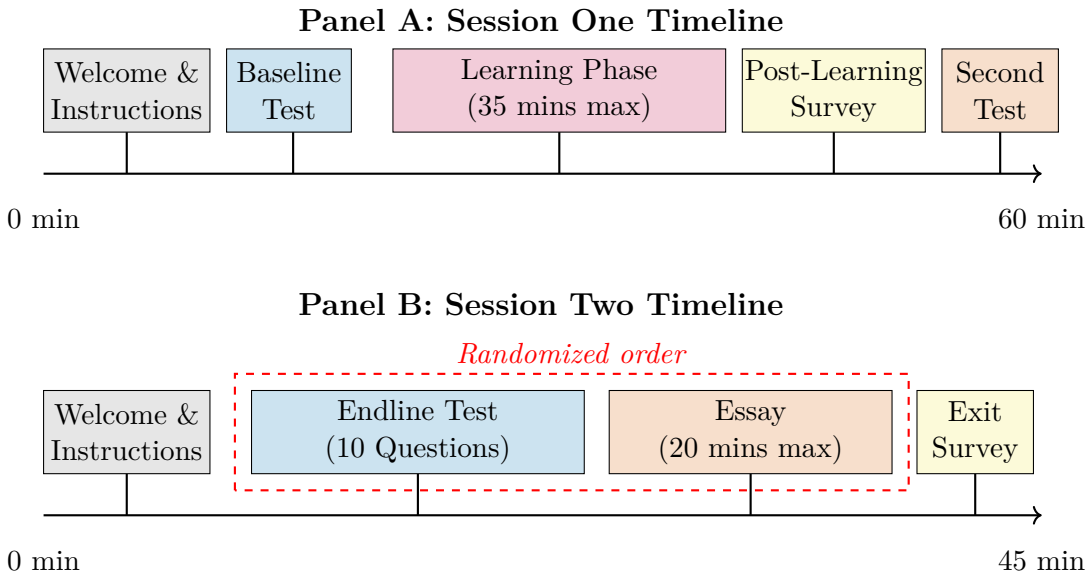
Our findings contribute to evidence that AI's educational impact depends on implementation details and usage patterns. In workplace settings, Brynjolfsson et al. (2025) find that AI helps novice call-center workers learn more effectively, while Budzyń et al. (2025) show that AI assistance during colonoscopies leads to skill deterioration. In educational settings, Bastani et al. (2025) find that unrestricted AI access harms learning in high school mathematics, but blocking solution-seeking while permitting explanation-seeking mitigates negative effects. Lehmann et al. (2024) document that students who copy-paste AI code learn less, but those requesting explanations learn more. Our results align with this pattern: AI enhances learning when students use it for conceptual understanding rather than automation.

These findings have implications for educational policy. Blanket policies prohibiting or permitting AI may be suboptimal. Our results suggest that approaches guiding students toward productive uses (concept explanation, feedback) while discouraging complete automation of cognitive tasks may be more effective. AI may help reduce achievement gaps by

assisting struggling students in reaching competency thresholds. The disconnect between objective learning gains and subjective enjoyment creates potential misalignment between student preferences and educational effectiveness, particularly for institutions relying on satisfaction metrics.

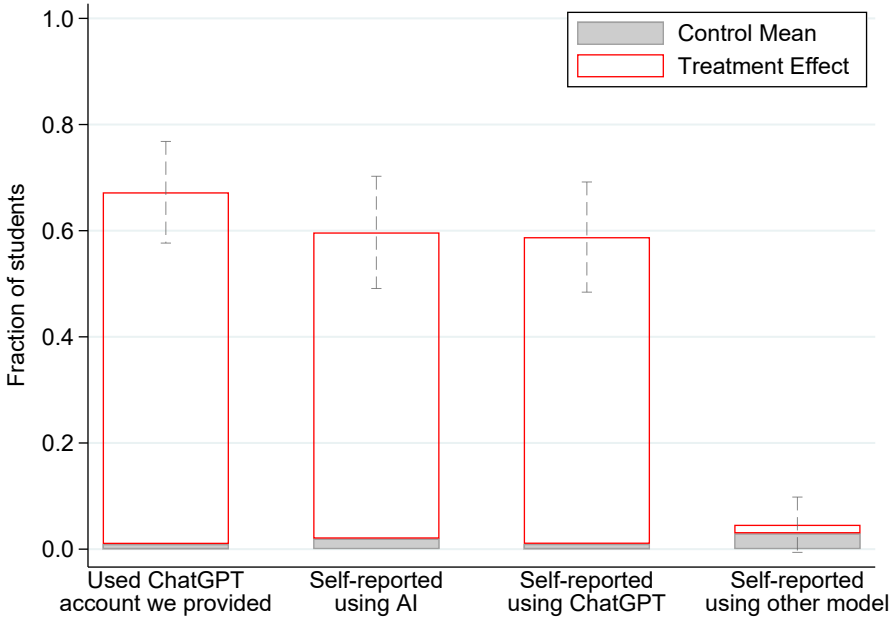
Figures and Tables

Figure 1: Experimental Sessions Timelines



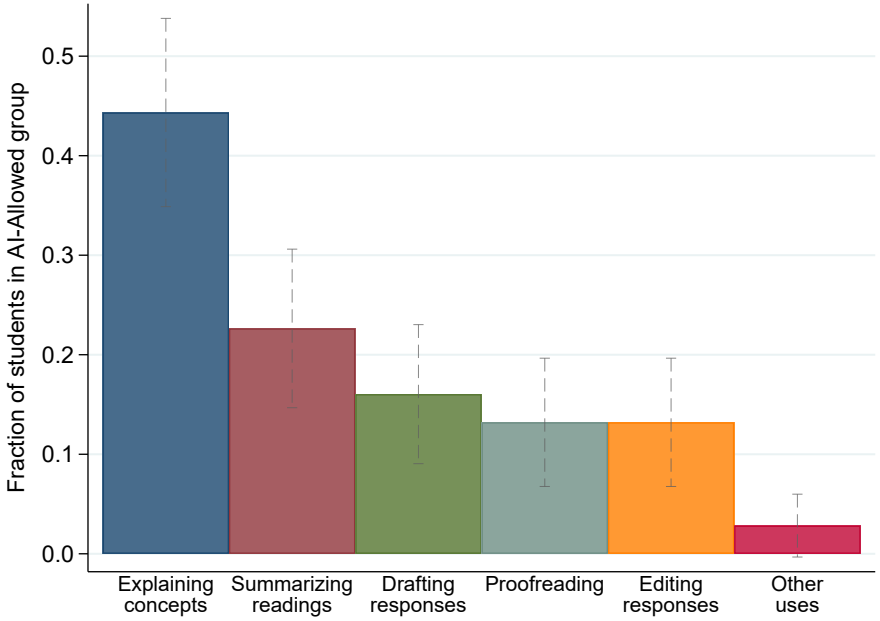
Notes: Panel A shows the structure of Session One, during which participants were randomly assigned to AI-allowed or AI-forbidden conditions during the Learning Phase. Panel B shows Session Two, conducted approximately one week later (mean = 6.9 days), during which all participants completed tasks without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

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



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

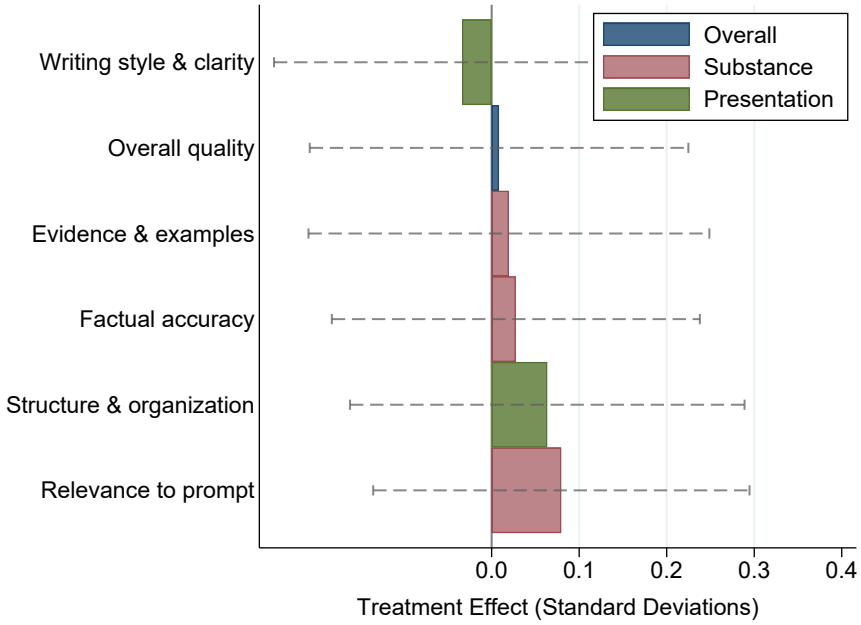
Figure 3: Types of Generative AI Use During the Learning Phase



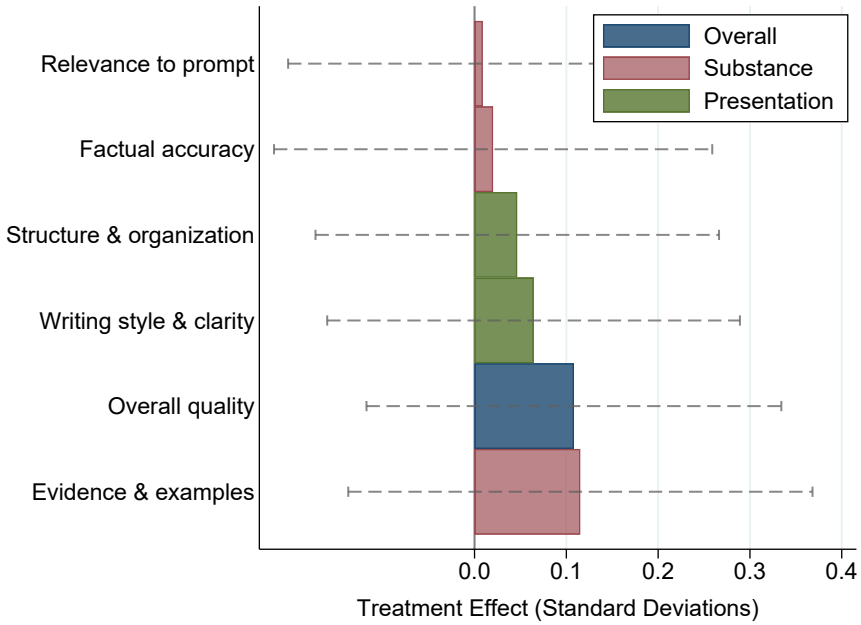
Notes: This figure shows the distribution of self-reported AI usage types among participants in the AI-allowed condition who used generative AI during Session One. Participants could select multiple usage types. The bars represent the proportion of all reported AI uses falling into each category: explaining concepts, summarizing readings, drafting responses, editing responses, proofreading, and other uses.

Figure 4: Effects of AI Access on Essay Quality

Panel A. Session One



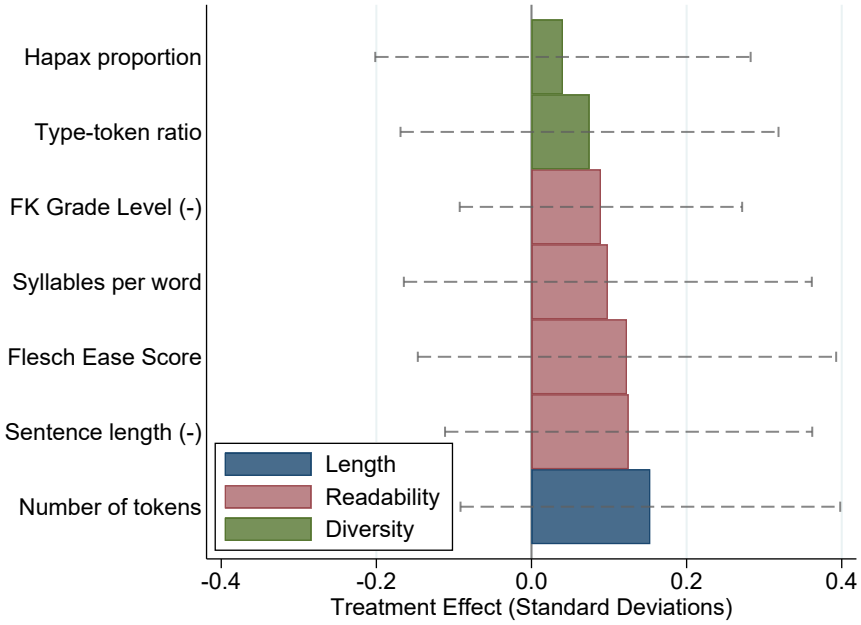
Panel B. Session Two



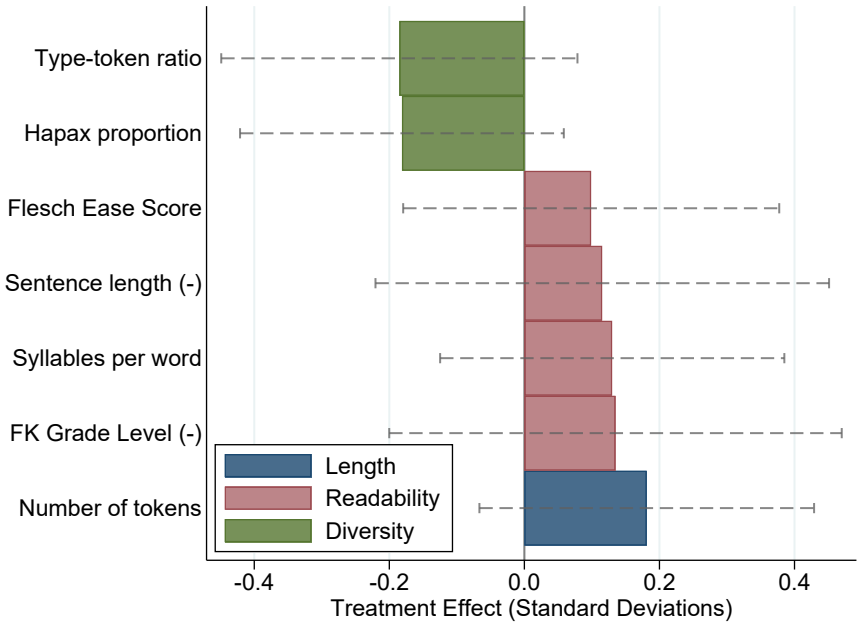
Notes: This figure presents treatment effects of AI access on essay quality dimensions, measured in standard deviations. Panel A shows effects on essays written during the learning phase in Session One. Panel B shows effects on essays written in Session Two (without AI access). Essays were evaluated by independent graders on six dimensions using a standardized rubric. Navy bars represent overall quality (grader's holistic assessment). Red bars represent substantive dimensions (factual accuracy, use of evidence and examples, and relevance to prompt). Green bars represent presentation dimensions (structure and organization, and writing style and clarity). Each dimension is scored on a scale from 1 to 10 and standardized to have mean 0 and standard deviation 1 in the control group. Horizontal lines represent 95 percent confidence intervals with standard errors clustered at the individual level.

Figure 5: Effects of AI Access on Essay Characteristics

Panel A. Session One



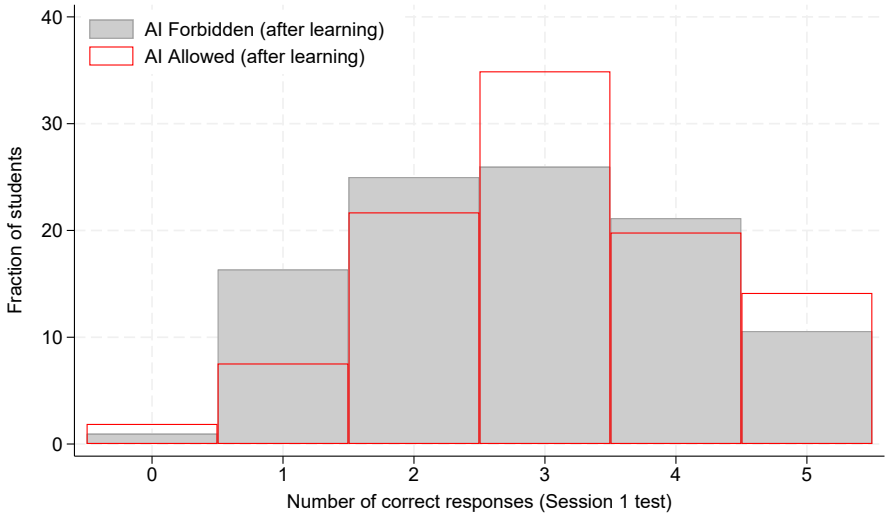
Panel B. Session Two



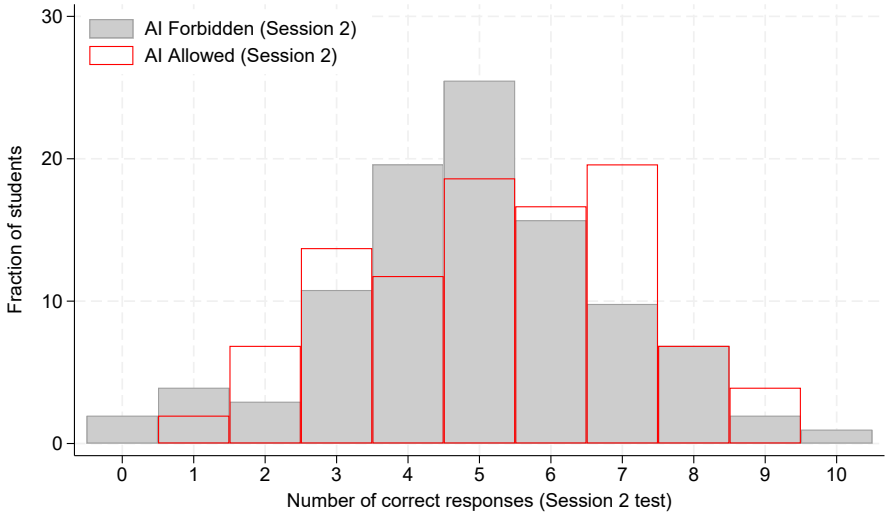
Notes: This figure presents treatment effects of AI access on quantitative characteristics of essays, measured in standard deviations. Panel A shows effects on essays written during the learning phase in Session One. Panel B shows effects on essays written in Session Two (without AI access). Blue bars represent measures of length (number of tokens), red bars represent measures of readability (syllables per word, Flesch Reading Ease score, sentence length, and Flesch-Kincaid grade level), green bars represent measures of lexical diversity (type-token ratio and hapax proportion), and orange bars represent measures of textual homogeneity (within-group cosine similarity and source text cosine similarity). Horizontal lines represent 95 percent confidence intervals with standard errors clustered at the individual level.

Figure 6: Distribution of Test Performance by Treatment Group

Panel A. Session One (Knowledge Acquisition)

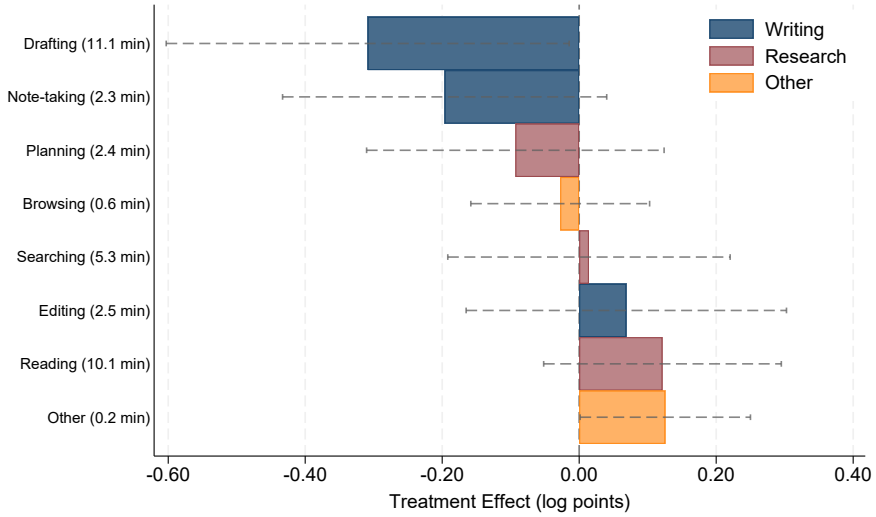


Panel B. Session Two (Knowledge Retention)



Notes: This figure shows the distribution of test scores by treatment assignment. The x-axis represents the number of correct responses (0-5) on the knowledge test. The y-axis shows the fraction of students achieving each score. Gray bars represent students in the AI Forbidden group, while red outlined bars represent students in the AI Allowed group. Panel A shows the distribution of scores on the immediate test administered at the end of Session One. Panel B shows the distribution of scores on the retention test administered in Session Two, approximately one week after the learning phase.

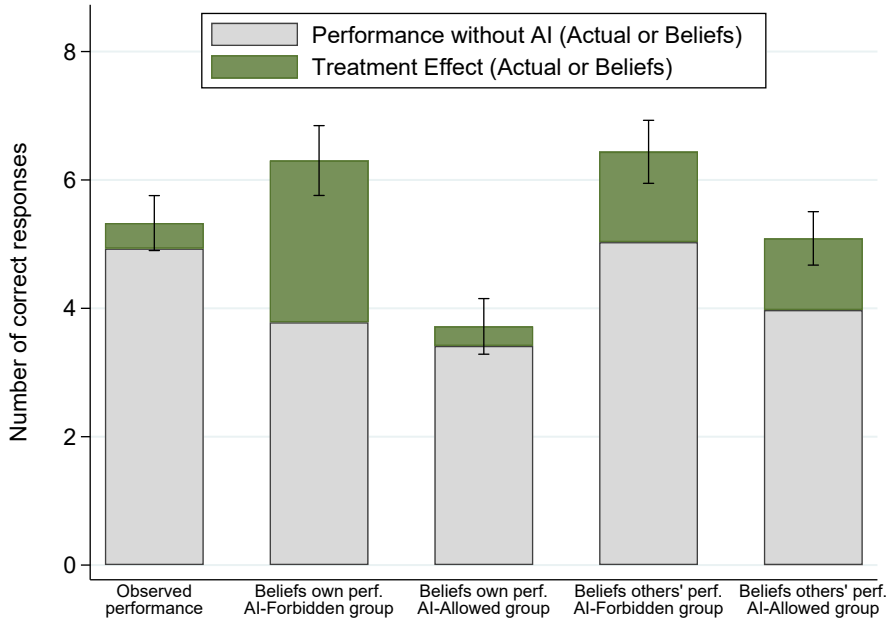
Figure 7: Mechanisms: Time Allocation Across Learning Activities



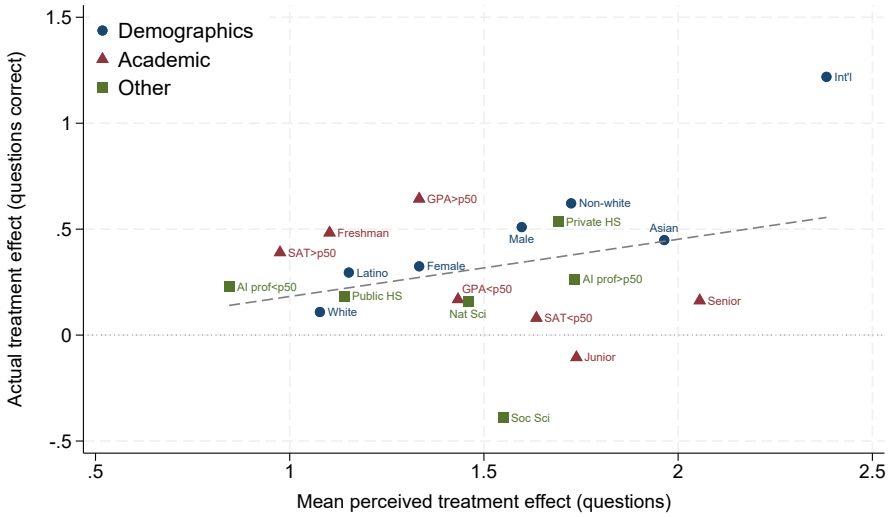
Notes: This figure presents treatment effects of AI access on self-reported time allocation across different learning activities, measured in log minutes. Activities are: searching for information online, reading source materials, taking notes, planning the essay, drafting the essay, editing the essay, browsing unrelated content, and other activities. Bars represent the estimated treatment effect for each activity, with horizontal lines showing 95 percent confidence intervals. Numbers in parenthesis indicate the mean in the control group.

Figure 8: Actual and Perceived Treatment Effects on Test Performance

Panel A. Beliefs vs. Actual Effects



Panel B. Perceived vs. Actual Effects by Subgroup



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

Table 1: Summary Statistics

	Completed Intake (1)	Attended Session One (2)	Attended Both Sessions (3)
Panel A. Demographic characteristics			
Age	20.211	20.148	20.132
Male	0.359	0.352	0.353
White	0.496	0.495	0.510
International student	0.270	0.267	0.275
Public high school	0.539	0.538	0.534
Panel B. Academic background			
Took SAT/ACT	0.742	0.762	0.770
SAT score (conditional)	1387.593	1385.566	1386.571
College GPA	3.680	3.684	3.686
Freshman	0.359	0.376	0.382
Already declared major	0.780	0.779	0.777
Natural-science major	0.297	0.305	0.309
Social-science major	0.363	0.357	0.353
Humanities/Arts/Lang. major	0.105	0.100	0.098
Study hours per week	16.878	16.495	16.623
Panel C. Attendance and comprehension			
Attended session 1	0.824	1.000	1.000
Attended session 2	0.797	0.971	1.000
Attended session 2 (cond. S1)	0.971	0.971	1.000
Attended both sessions	0.797	0.971	1.000
Days between sessions	6.943	6.976	6.971
Comprehension score (0-5)	4.768	4.767	4.765
Panel D. Experimental conditions			
Assigned Blockchain topic	0.365	0.367	0.368
Assigned CRISPR topic	0.327	0.329	0.324
Assigned Carbon-capture topic	0.308	0.305	0.309
Baseline self-assessment (0-10)	1.517	1.524	1.520
Fraction correct (baseline)	0.303	0.304	0.310
Shared info between sessions	0.064	0.064	0.064
Looked up info about topic	0.049	0.049	0.049
Studied topic for session 2	0.005	0.005	0.005
Number of observations	256	210	204

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

Table 2: Balance of Baseline Characteristics by Treatment Assignment

	Treatment group:		Difference:	
	AI Forbidden	AI Allowed	$\hat{\beta}$	(SE)
Panel A. Demographic characteristics				
Age	20.016	20.400	0.378	(0.225)*
Male	0.405	0.315	-0.090	(0.060)
White	0.524	0.469	-0.059	(0.063)
International student	0.286	0.254	-0.030	(0.055)
Public high school	0.540	0.538	-0.004	(0.063)
Panel B. Academic background				
Took SAT/ACT	0.778	0.708	-0.070	(0.054)
SAT score (conditional)	1392.371	1382.554	-8.823	(21.080)
College GPA	3.694	3.666	-0.025	(0.038)
Freshman	0.357	0.362	0.010	(0.061)
Already declared major	0.825	0.734	-0.093	(0.053)*
Natural Sciences major	0.294	0.300	0.009	(0.058)
Social Sciences major	0.405	0.323	-0.089	(0.061)
Humanities/Arts/Languages major	0.111	0.100	-0.008	(0.038)
Study hours per week	16.864	16.892	-0.056	(1.194)
Panel C. Attendance and comprehension				
Attended session 1	0.825	0.822	-0.007	(0.048)
Attended session 2	0.810	0.785	-0.022	(0.050)
Attended session 2 (cond. S1)	0.981	0.962	-0.018	(0.022)
Attended both sessions	0.810	0.785	-0.022	(0.050)
Days between sessions	6.975	6.913	-0.045	(0.155)
Comprehension score (0-5)	4.769	4.766	-0.010	(0.065)
Panel D. Experimental conditions				
Assigned Blockchain topic	0.375	0.355	-0.020	(0.066)
Assigned CRISPR topic	0.327	0.327	0.005	(0.065)
Assigned Carbon-capture topic	0.298	0.318	0.015	(0.062)
Baseline self-assessed knowledge (0-10)	1.558	1.477	-0.071	(0.252)
Fraction correct (baseline)	0.306	0.301	-0.006	(0.038)
N (Students)	126	130	256	

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

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

Outcome	Control mean (1)	Effect (2)
Used ChatGPT (revealed)	0.010	0.663***
		(0.049)
No. prompts	0.000	0.762***
		(0.060)
Big pasting event	0.154	0.100*
		(0.057)
Used any AI (self-reported)	0.020	0.599***
		(0.051)
Used ChatGPT (self-reported)	0.010	0.602***
		(0.050)
Used other AI	0.030	0.009
		(0.026)
N	104	210

Notes: Each row reports the effect of being assigned to the *AI-allowed* group on the indicated measure of AI usage. *Used ChatGPT (revealed)* equals one if at least one prompt was recorded in the ChatGPT account provided to the student or if a proctor directly observed the student using ChatGPT. *No. prompts* is the total number of prompts sent to the provided ChatGPT account during the session. *Big pasting event* equals one if the student's essay grew by more than three standard deviations above the mean interval change (approximately 715 characters) in any single 30-second autosave window. For reference, an average typist (approximately 40 words per minute) would need 3–4 minutes to produce this volume of text; a fast typist (70–80 WPM) could do it in under 2 minutes. The remaining rows present 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: Effects of Generative AI Access on Essay Quality

Outcome	Session One			Session Two		
	Control mean (1)	ITT (2)	TOT (3)	Control mean (4)	ITT (5)	TOT (6)
Panel A. Overall quality						
Overall quality (human)	6.675	−0.047	−0.070	5.685	0.235	0.342
		(0.236)	(0.354)		(0.262)	(0.389)
Overall quality (AI)	3.941	0.168	0.244	3.000	0.262	0.379
		(0.200)	(0.288)		(0.178)	(0.258)
Quality index (human)	6.618	0.069	0.103	5.650	0.134	0.194
		(0.204)	(0.307)		(0.237)	(0.347)
Quality index (AI)	4.673	0.104	0.150	3.571	0.256	0.366
		(0.177)	(0.254)		(0.167)	(0.245)
Panel B. Individual dimensions (human and AI averaged)						
Accuracy	5.887	0.107	0.155	4.699	0.169	0.248
		(0.171)	(0.247)		(0.159)	(0.236)
Evidence	5.353	0.015	0.022	3.626	0.303*	0.442*
		(0.194)	(0.279)		(0.170)	(0.254)
Relevance	5.670	0.011	0.016	4.965	0.186	0.261
		(0.170)	(0.245)		(0.158)	(0.225)
Organization	5.350	0.305	0.444	4.738	0.238	0.339
		(0.203)	(0.295)		(0.196)	(0.282)
Writing style	5.860	0.305	0.440	5.146	0.298*	0.435*
		(0.198)	(0.286)		(0.162)	(0.240)
N	104	210	210	102	204	204

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

Table 5: Effects of Generative AI Access on Essay Characteristics

	Session One			Session Two		
Outcome	Control mean (1)	ITT (2)	TOT (3)	Control mean (4)	ITT (5)	TOT (6)
Length	0.000	0.153	0.230	0.000	0.187	0.275
		(0.125)	(0.187)		(0.129)	(0.191)
Readability index	0.000	0.093	0.136	0.000	0.048	0.070
		(0.084)	(0.123)		(0.084)	(0.122)
Lex. diversity index	0.000	0.080	0.119	0.000	−0.194*	−0.284
		(0.113)	(0.169)		(0.117)	(0.172)
Within-group sim.	0.790	−0.005	−0.007	0.761	−0.014	−0.021
		(0.007)	(0.010)		(0.015)	(0.022)
Source text sim.	0.752	0.006	0.009	0.715	−0.011	−0.016
		(0.013)	(0.021)		(0.021)	(0.032)
AI-detected (frac.)	0.128	0.106**	0.156**	0.041	−0.000	−0.000
		(0.050)	(0.072)		(0.028)	(0.041)
N	104	210	210	102	204	204

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

Table 6: The Impact of Generative AI on Learning

	Session One			Session Two		
Outcome	Control mean (1)	ITT (2)	TOT (3)	Control mean (4)	ITT (5)	TOT (6)
Self-assessed	4.788	−0.017	−0.024	3.842	0.186	0.265
		(0.193)	(0.277)		(0.190)	(0.270)
Frac. correct	0.563	0.064**	0.092**	0.493	0.037*	0.053*
		(0.031)	(0.045)		(0.022)	(0.031)
Test score (SD)	0.000	0.251**	0.364**	0.000	0.204*	0.294*
		(0.122)	(0.177)		(0.116)	(0.168)
N	104	210	210	102	204	204

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

Table 7: The Distributional Impact of Generative AI on Learning

	Session One			Session Two		
Outcome	Control mean (1)	ITT (2)	TOT (3)	Control mean (4)	ITT (5)	TOT (6)
Above 20th pctile	0.990	−0.010	−0.014	0.941	0.040	0.057
		(0.016)	(0.023)		(0.028)	(0.040)
Above 40th pctile	0.827	0.079	0.114	0.804	−0.002	−0.003
		(0.048)	(0.070)		(0.050)	(0.072)
Above 60th pctile	0.577	0.118*	0.166*	0.353	0.077	0.112
		(0.065)	(0.091)		(0.064)	(0.093)
Above 80th pctile	0.317	0.051	0.074	0.098	0.012	0.017
		(0.059)	(0.085)		(0.043)	(0.060)
Above 100th pctile	0.106	0.029	0.042	0.010	−0.009	−0.013
		(0.043)	(0.063)		(0.009)	(0.013)
N	104	210	210	102	204	204

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

Table 8: Mechanisms: Engagement, Experience, and Academic Integrity

Outcome	Control mean (1)	ITT (2)	TOT (3)
Panel A. Engagement			
Opened PDF	0.712	0.065	0.094
		(0.060)	(0.087)
Time on task (Qualtrics, min.)	32.611	−0.223	−0.324
		(0.680)	(0.988)
Time on task (self-reported, min.)	34.312	−0.551	−0.810
		(0.671)	(0.986)
Remembers topic	0.960	0.000	0.000
		(0.028)	(0.040)
Panel B. Learning experience			
Enjoyed learning (continuous)	5.221	0.597*	0.876*
		(0.309)	(0.462)
Enjoyed learning (above median)	0.337	0.148**	0.216**
		(0.064)	(0.094)
Found effective (continuous)	5.000	0.038	0.055
		(0.287)	(0.411)
Found effective (above median)	0.423	0.045	0.065
		(0.067)	(0.098)
Panel C. Academic integrity			
Caught by proctor	0.029	0.064**	0.094**
		(0.032)	(0.046)
Self-reported AI use	0.020	0.086**	0.124**
		(0.034)	(0.049)
Any integrity violation	0.050	0.126***	0.182***
		(0.044)	(0.064)
Both violations	0.000	0.019	0.027
		(0.013)	(0.019)
N	104	210	210

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

References

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

Autor, D. and Thompson, N. (2025). Expertise. *Journal of the European Economic Association*, page jvaf023.

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.

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.

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

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

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

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

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

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

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

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

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

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

Cui, Z. K., Demirer, M., Jaffe, S., Musolff, L., Peng, S., and Salz, T. (2024). The Effects of Generative AI on High Skilled Work: Evidence from Three Field Experiments with Software Developers.

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

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.

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

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

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

Herrman, J. (2025). Everyone is cheating their way through college: ChatGPT has unraveled the entire academic project. *New York Magazine*.

Hsu, H. (2025). What happens after A.I. destroys college writing? *The New Yorker*.

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

Lehmann, M., Cornelius, P. B., and Sting, F. J. (2024). AI Meets the Classroom: When Does ChatGPT 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.

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.

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

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

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

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

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

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.

Yakura, H., Namikoshi, K., and Shinoda, K. (2024). Empirical evidence of large language model's influence on human spoken communication. *arXiv preprint arXiv:2409.01754*.

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

Appendix

A Appendix Figures and Tables

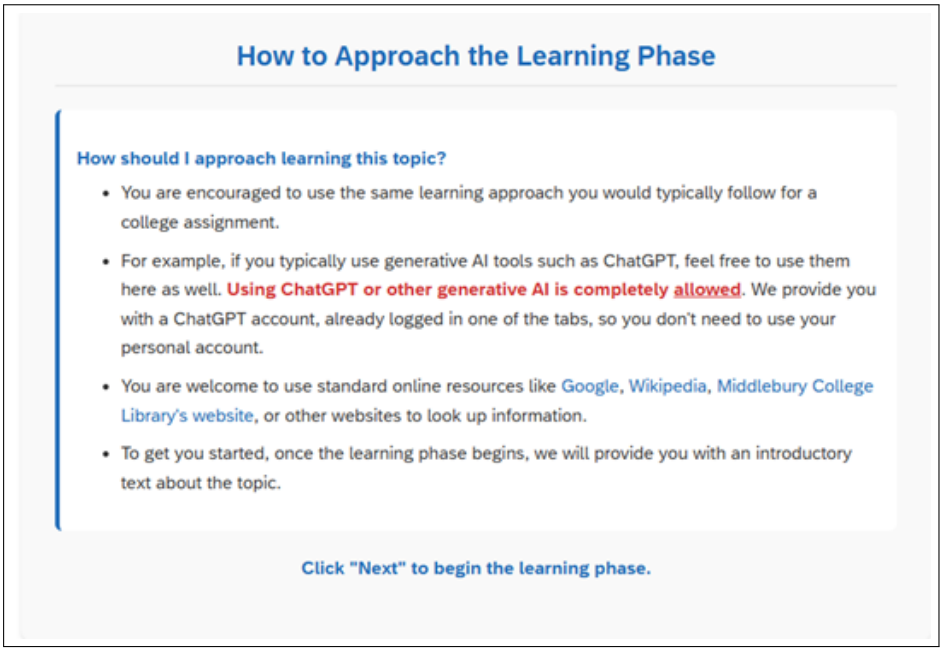
Figure A1: Computer Lab Setup with Privacy Dividers



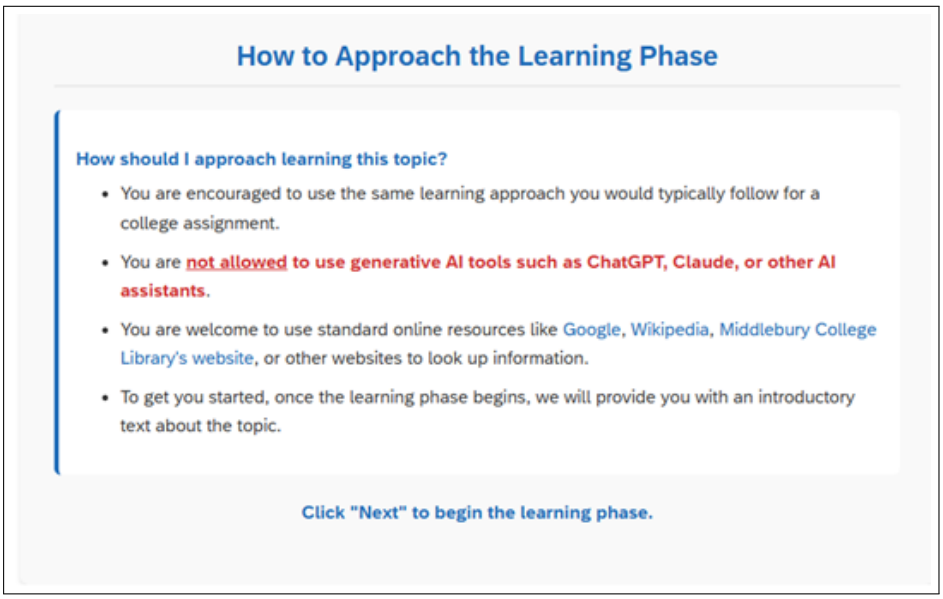
Note: This figure shows the computer lab setup used during experimental sessions. Each workstation was equipped with privacy dividers to minimize distractions and prevent participants from viewing other screens.

Figure A2: Treatment Instructions Displayed to Participants

Panel A. AI-Allowed Condition



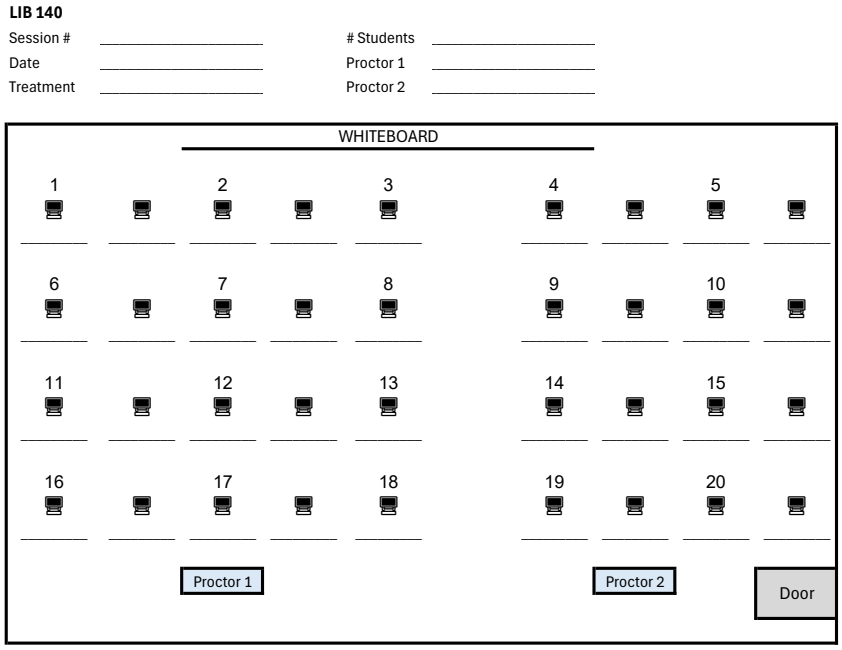
Panel B. AI-Forbidden Condition



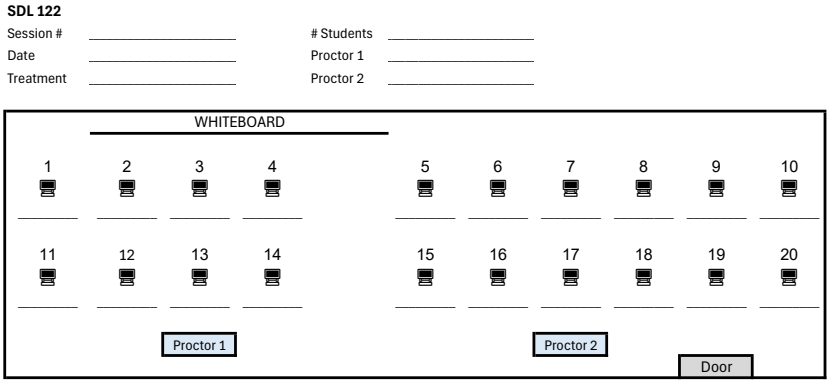
Note: This figure shows the treatment-specific instructions displayed to participants in the survey interface. Panel A shows the message for participants randomly assigned to the AI-allowed condition, who could use generative AI tools during the 35-minute learning phase. Panel B shows the message for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

Figure A3: Laboratory Seating Charts

Panel A. Library Computer Lab (LIB 140)



Panel B. Language Building Computer Lab (SDL 122)



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

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	0.119	0.113	0.073	0.060
	(0.104)	(0.110)	(0.104)	(0.103)
White	−0.180*	−0.065	−0.177	0.010
	(0.108)	(0.112)	(0.114)	(0.129)
International student	0.148	0.093	0.168	0.044
	(0.117)	(0.127)	(0.147)	(0.160)
Private high school	−0.019	−0.083	−0.148	−0.142
	(0.096)	(0.106)	(0.115)	(0.120)
Panel B. Academic background				
Natural Sciences major	−0.119	−0.167	−0.162	−0.089
	(0.099)	(0.106)	(0.132)	(0.124)
Social Sciences major	0.112	0.258**	−0.006	0.264**
	(0.098)	(0.101)	(0.131)	(0.128)
High GPA (above median)	−0.144	−0.114	−0.060	−0.052
	(0.090)	(0.099)	(0.106)	(0.122)
Freshman	−0.129	−0.071	−0.234*	−0.014
	(0.098)	(0.109)	(0.137)	(0.140)
Sophomore	0.118	0.119	−0.073	0.043
	(0.101)	(0.125)	(0.138)	(0.178)
Junior	−0.037	−0.016	−0.075	−0.098
	(0.126)	(0.133)	(0.158)	(0.158)
Panel C. AI experience and beliefs				
Frequent AI user (above median)	0.284***	0.251**	0.273***	0.173
	(0.088)	(0.098)	(0.100)	(0.114)
Early AI adopter (pre−Fall 2024)	0.151	0.161	−0.000	0.021
	(0.095)	(0.103)	(0.112)	(0.130)
High AI proficiency (above median)	0.183**	0.208**	0.011	0.130
	(0.088)	(0.101)	(0.108)	(0.124)
High perceived AI effect (above median)	0.124	0.234**	0.044	0.208*
	(0.094)	(0.101)	(0.095)	(0.105)
N (Students)	106	102	101	100

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

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

	Session One			Session Two		
Outcome	Control mean (1)	ITT (2)	TOT (3)	Control mean (4)	ITT (5)	TOT (6)
Panel A. Length						
Essay is empty	0.000	0.009	0.013	0.030	0.009	0.013
		(0.009)	(0.013)		(0.017)	(0.025)
Tokens	202.913	15.115	22.656	150.724	12.762	18.762
		(12.318)	(18.430)		(8.809)	(13.042)
Panel B. Readability						
Sentence length	−27.104	4.608	6.712	−25.545	3.210	4.721
		(4.604)	(6.719)		(4.785)	(7.041)
Syllables/word	1.775	0.015	0.022	1.714	0.015	0.021
		(0.016)	(0.023)		(0.017)	(0.024)
FK grade level	−15.921	1.674	2.438	−14.598	0.622	0.903
		(1.788)	(2.609)		(1.284)	(1.864)
Flesch ease	29.193	3.795	5.528	35.901	1.284	1.864
		(4.752)	(6.937)		(3.410)	(4.950)
Panel C. Lexical diversity						
TTR	0.723	0.009	0.014	0.718	−0.018	−0.026
		(0.010)	(0.015)		(0.011)	(0.017)
Hapax prop.	0.579	0.005	0.007	0.566	−0.017	−0.025
		(0.014)	(0.020)		(0.016)	(0.023)
Panel D. Textual similarity						
Cosine sim.	0.790	−0.005	−0.007	0.761	−0.016	−0.024
		(0.007)	(0.010)		(0.015)	(0.023)
N	104	210	210	102	204	204
Source text sim.	0.752	0.006	0.009	0.715	−0.011	−0.016
		(0.013)	(0.021)		(0.021)	(0.032)

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

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

	Session One			Session Two		
Outcome	Control mean (1)	ITT (2)	TOT (3)	Control mean (4)	ITT (5)	TOT (6)
Score ≥ 1	1.000	−0.007	−0.010	0.994	−0.007	−0.010
		(0.005)	(0.008)		(0.009)	(0.013)
Score ≥ 3	0.978	−0.026	−0.039	0.913	−0.032	−0.046
		(0.018)	(0.026)		(0.035)	(0.051)
Score ≥ 5	0.880	−0.035	−0.052	0.749	0.001	0.002
		(0.040)	(0.060)		(0.052)	(0.076)
Score ≥ 7	0.577	0.013	0.019	0.364	0.061	0.087
		(0.057)	(0.084)		(0.046)	(0.066)
Score ≥ 9	0.161	0.011	0.016	0.075	0.055**	0.081**
		(0.041)	(0.060)		(0.025)	(0.038)
N	392	791	791	376	757	757

Notes: Each row reports the effect of being assigned to the *AI-allowed* group on the probability of achieving at least the indicated essay quality threshold. Quality scores range from 0 (lowest) to 10 (highest). The unit of observation is the individual grader-essay pair. 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 grader and batch fixed effects and controls selected through a double-lasso procedure (Belloni et al., 2014). Standard errors clustered at the student level in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

B Empirical Appendix

B.1 Reading Materials

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

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

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

Table B1: Readability Metrics for Topic Texts

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

B.2 Writing Assessment Prompts

This appendix presents the analytical writing prompts developed for each of the three topics. For each topic, we designed two complementary prompts of comparable complexity, one for use in Session One and the other in Session Two. These prompts were carefully constructed to assess participants' analytical reasoning abilities.

Each prompt requires participants to analyze relationships between concepts, evaluate the relative importance of different factors, and support their analysis with specific examples from the reading materials. The prompts were designed to be sufficiently challenging to differentiate between levels of understanding while remaining accessible to participants with no prior knowledge of the topics beyond the provided reading materials.

Blockchain Technology Prompts

- **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 Prompts

- **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 Prompts

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

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

Graders evaluated each essay on six dimensions using a 0–10 scale. *Accuracy of Content* measures the correctness and accuracy of the information presented in the essay. *Use of Evidence and Examples* assesses the integration of specific evidence, examples, and data from the reading to support arguments. *Relevance to the Prompt* evaluates how well the essay addresses the specific prompt and stays on topic throughout. *Organization and Structure* measures the logical structure, coherence, flow, and effectiveness of transitions. *Writing Style and Clarity* assesses clarity, readability, grammar, and precision of language. *Overall Essay Quality* captures the grader's holistic assessment of the essay. Graders also

reported their prior familiarity with each of the three topics (blockchain, carbon capture, and CRISPR) on a 0–10 scale before beginning the grading task.

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

B.4 AI Essay Grading

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

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

The system prompt is as follows:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table B2: 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	0.432	0.683	3.895	0.013	0.018
		(0.277)	(0.438)		(0.378)	(0.530)
Frac. correct	0.528	0.054	0.084	0.584	0.013	0.018
		(0.053)	(0.083)		(0.028)	(0.039)
Test score (SD)	−0.139	0.236	0.369	0.469	0.136	0.212
		(0.205)	(0.317)		(0.193)	(0.297)
N	39	77	77	38	75	75
Panel B. CRISPR						
Self-assessed	4.735	−0.304	−0.410	3.606	0.380	0.496
		(0.368)	(0.502)		(0.302)	(0.388)
Frac. correct	0.647	0.087	0.117	0.470	0.073	0.095
		(0.062)	(0.084)		(0.059)	(0.080)
Test score (SD)	0.330	0.344	0.460	−0.121	0.375	0.488
		(0.245)	(0.329)		(0.301)	(0.413)
N	34	69	69	33	66	66
Panel C. Carbon capture						
Self-assessed	4.839	0.111	0.174	4.033	0.088	0.134
		(0.459)	(0.722)		(0.437)	(0.671)
Frac. correct	0.516	0.098	0.147	0.420	0.016	0.024
		(0.062)	(0.095)		(0.039)	(0.058)
Test score (SD)	−0.187	0.386	0.579	−0.377	0.081	0.125
		(0.246)	(0.374)		(0.199)	(0.300)
N	31	64	64	31	63	63

Notes: This table replicates the main test score results (Table 6) separately by assigned topic. Each panel restricts the sample to students assigned the indicated topic. All specifications mirror those in the main table: ITT estimates from OLS with double-lasso-selected controls (Belloni et al., 2014) and strata fixed effects; TOT estimates instrument ChatGPT use with random assignment. Standard errors clustered at the student level in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.