# The Effects of Course-Integrated AI Tutoring on Student Performance and Engagement: A Randomized University Trial *

Jing Liu, Tracy Sweet, Marty Haoyuan Chen,
Jonathan Engelberg, Martyn Clark, Axel Persaud, Alia Lancaster,
Jeffrey Hollingsworth, Jennifer King Rice, Megan Masters

University of Maryland, College Park

September 28, 2026

**Abstract**

We evaluate access to a course-integrated generative AI (GenAI) tutor in a randomized controlled trial involving 2,379 undergraduate students and 30 instructors across multiple disciplines in a large U.S. public university. Among sections of the same course, tutor access reduced final grades by 0.37 standard deviations and learning management system participation by 0.90 standard deviations. Participation declines were similarly large in the full sample, and estimated academic losses were greater among first-generation students. These findings suggest that introducing an AI tutor may displace engagement with existing learning activities without improving academic performance, with potentially differential impact across student groups.

**JEL Codes: I23, I21, C93, O33.**
**Keywords: Generative AI; Higher Education; AI Tutoring; Randomized Controlled Trial; Academic Performance; Student Engagement.**

*The authors would like to thank Heather Hill, Xiaoyang Ye, Yang Song for their feedback on earlier drafts. Correspondence regarding the paper should be sent to Jing Liu (jliu28@umd.edu).

# 1 Introduction

Generative artificial intelligence (GenAI) is reshaping human productivity—not only by changing how people perform tasks, but also by altering the learning procedures through which they acquire the knowledge and skills to perform them (Deng et al., 2025; Kasneci et al., 2023; McGrath et al., 2025; Purwono et al., 2025). In education, GenAI offers the prospect of personalized instructional support at unprecedented scale (Farrokhnia et al., 2024; Grassini, 2023; Kasneci et al., 2023), while also raising a central challenge: how to design and deploy GenAI-powered tools that enhance rather than replace human learning. Although intelligent tutoring systems have a long history in educational research and practice, the widespread availability of large language model (LLM)–based conversational tools since late 2022 has accelerated efforts to integrate AI into everyday learning environments (Winkler and Söllner, 2018; Xu et al., 2024; Al-Abdullatif et al., 2023). By providing on-demand explanations, feedback, and practice tailored to students' needs and learning stages, these tools could expand access to individualized tutoring and improve learning outcomes (Michel-Villarreal et al., 2023; Rudolph et al., 2023).

However, focusing on efficiency of performance with AI assistance is not the same as developing the knowledge and skills needed to perform independently. AI may support the cognitive work through which students develop understanding, or potentially serve as a complete substitute for it. Recent studies illustrate this tension: access to general-purpose AI can improve performance on assisted tasks while undermining subsequent independent performance—the so-called "learning penalty"—although these adverse effects may be mitigated by purpose-built tutoring safeguards (Strömberg et al., 2026; Bastani et al., 2025). AI assistance can reduce students' mental effort or cognitive load (Avello-Martínez et al., 2024; Deng et al., 2025; Wu et al., 2024), potentially easing barriers to understanding. Yet these reductions may come at a cost: Stadler et al. (2024) found that LLM assistance reduced mental effort while compromising the depth of scientific inquiry. An EEG-based essay-writing study likewise documented differences in neu-

ral engagement across AI-assisted and independent writing conditions (Kosmyna et al., 2025). Whether embedding GenAI in routine educational settings strengthens learning or displaces the effort and engagement that sustain it therefore remains a consequential empirical question.

In higher education, the adoption of GenAI has outpaced evidence on its educational consequences, raising concerns about academic integrity and equitable access to learning support (Deng et al., 2025; Purwono et al., 2025; Walczak and Cellary, 2023; Cotton et al., 2024). Beyond facilitating academic misconduct, GenAI may reshape how students engage with the broader learning environment. Its availability could change when students study, whether they consult primary course materials, and how often they seek assistance from instructors or peers. These behavioral consequences remain underexplored, yet they may be central to understanding whether AI complements or displaces existing learning activities. Moreover, unequal access to AI tools and differences in how students use them may reinforce educational inequalities (Yu et al., 2024). Together, these concerns underscore the need for systematic evidence on both academic and behavioral outcomes to guide institutional policy and the responsible adoption of GenAI in higher education.

The current study presents findings from, to our knowledge, the first large-scale randomized experiment evaluating the effects of AI tutoring on student learning at a large U.S. public university. Our research setting and design offer several strengths. First, the study includes approximately 2,400 undergraduate students and 30 instructors across courses spanning STEM, the social sciences, the humanities, and education. This disciplinary breadth, together with students' diverse demographics and socioeconomic backgrounds, strengthens the potential generalizability of our findings. Second, we randomized instructors within blocks of identical or similar courses, allowing comparisons within broadly comparable instructional settings. We report estimates for the full sample and an exact-match subsample restricted to blocks containing sections of the same course. Third, the AI tutor we examine is a GPT-4o–based virtual study assistant (VSA) that uses

retrieval-augmented generation (RAG) to draw on instructor-selected course materials within the institution's Enterprise Learning Management System (ELMS). Instructors select the course materials made available to the tutor, choose between direct instruction and tutoring modes, and can adjust their instruction based on the types of prompts or questions students are generating when using the VSA. This instructor-guided configuration and course site integration allow us to evaluate AI tutoring within an applied routine learning environment rather than in a stand-alone experimental setting. Finally, alongside course performance, we collect rich data on students' online course activity and interactions with the AI tutor, as well as survey responses from both students and instructors, allowing us to investigate potential mechanisms underlying the observed effects.

Conducted during the fall semester of the 2025–26 academic year, our study yields several important findings. Consistent with recent evidence (Bastani et al., 2025; Strömberg et al., 2026), we found that access to the AI tutor negatively affected student final course grades by 0.27–0.37 standard deviations (SDs), with the clearest evidence in the exact-match sample. Although several studies have found that GenAI improves short-term performance (Bastani et al., 2025; Strömberg et al., 2026), our estimated effects on homework scores are null or negative.

Alongside the adverse academic estimates, we found substantial reductions in students' engagement with the existing course platform. In the exact-match sample, AI tutor access reduced course page views ($-0.37$ SDs), active days ($-0.51$ SDs), and recorded participation in instructor-designed online activities ($-0.90$ SDs), with similar reductions in the full sample. Some substitution is expected given the tool's intended functions—for example, students may ask the tutor about course logistics rather than consult the course page. However, the decline in recorded participation, which includes discussion responses, assignment submissions, and quizzes, raises concerns about substitution beyond information retrieval. Survey responses provided complementary descriptive evidence: compared with nonusers in treatment classrooms, AI tutor users were more likely to re-

port asking instructors fewer questions about course content and sending the instructor fewer emails or ELMS messages, potentially changing the way in which these learners are engaged in the learning process.

We also found differences across student groups in both AI tutor use and its effects. Hispanic students were more likely to use the AI tutor than their White peers and Students in higher high school GPA quartiles (among study participants) generally had lower estimated uptake than those in the bottom quartile, although differences were not uniformly significant. Evidence of heterogeneity in the effects on learning and behavior was generally limited, however, the estimated final-grade losses among first-generation students were more than twice those among continuing-generation students in both samples, raising important concerns about potentially unequal impact of deploying AI tools.

Analysis of student requests provided additional descriptive evidence about how students used the tutor. Approximately 73.8% of requests sought information, explanations, or solutions, compared with 11.3% involving practice and test preparation and 0.7% seeking feedback on students' own work. While these categories do not directly provide evidence on cognitive offloading, the patterns we observe suggest that students predominantly used the VSA for passive, on-demand assistance rather than actively seeking evaluation of their own attempts.

These findings should be interpreted in light of the institutional and implementation context. Students in both conditions had free access to other university-supported AI tools, including ChatGPT and Gemini. Given that only approximately 15% of treated students used the VSA at least once, it is possible that instructors introducing and encouraging use of the course-specific tutor may have changed GenAI use beyond the VSA itself. Instructors' and researchers' ability to view student–tutor interactions may also have influenced what students were willing to ask, limiting comparability with privately used, stand-alone tools. Finally, few instructors switched from the default direct-instruction mode to the tutor mode, which guides students toward a correct response, rather than

providing them with a correct answer immediately (only two instructors ended the term with this mode in place). The findings reported here thus concern the VSA as configured and introduced in this study, rather than establishing the effects of more scaffolded AI tutoring.

Our study builds on the growing literature on GenAI in education, particularly the still-limited experimental evidence from higher education.[1] In K–12 settings, Sungu et al. (2026) find that instructor access to GenAI reduced students' intrinsic motivation without improving exam scores overall, with adverse academic effects among students of initially lower-performing instructors. Bastani et al. (2025) show that general-purpose AI access impaired subsequent independent performance in high school mathematics, whereas tutoring guardrails mitigated this harm. Using a difference-in-differences design and drawing on survey data, Strömberg et al. (2026) similarly find that GenAI adoption in high school reduced exam performance despite initially improving homework scores.

Recent experiments emphasize the importance of engagement and instructional structure. Oreopoulos and Low (2026) find modest achievement gains from Khan Academy combined with Khanmigo, but limited sustained tutor use. However, their design does not isolate AI's incremental contribution. Comparing AI support with the same practice platform without AI, Oreopoulos et al. (2026) find that the AI-enhanced platform leads to slower progression. They also find suggestive evidence of improved recovery from mistakes, and modest, marginally significant delayed-learning gains on one practice exercise under mastery-based progression. Together, these studies indicate that GenAI's educational effects depend on tutor configuration, instructional scaffolding, and realized engagement, and that improvements during assisted practice may not necessarily translate into independent learning.

In higher education, Barcaui (2025) uses an RCT to examine the impact of GenAI on undergraduates learning AI. The findings closely align with those from Bastani et al.

[1] For a recent review, see Wang and Fan (2025).

(2025) and Strömberg et al. (2026)—unrestricted use of ChatGPT contributes to cognitive-offloading, reduces student effort, and negatively impacts student knowledge retention. In contrast, Kestin et al. (2025) compare a customized AI tutor with an in-class active learning lesson in an undergraduate physics course at Harvard University. Students assigned to the tutor achieved substantially larger immediate learning gains in less time and reported greater engagement and motivation. These findings demonstrate the potential of purpose-built AI tutoring. However, its applicability beyond a single course at a highly selective university remains uncertain. The distinct findings from these two studies also point to the importance of evaluating purpose-built AI and general purpose AI learning tools separately. Lastly, Thoeni and Fryer (2026) evaluate a RAG-based chatbot in a semester-long RCT involving 500 students in a marketing course and find no statistically significant effects on achievement, interest, self-efficacy, or engagement. Their results suggest that grounding AI assistance in course-specific materials alone may be insufficient to improve educational outcomes.

Although informative, these studies leave several important questions unresolved. First, the higher-education trials examine individual courses, leaving limited evidence across disciplines and student populations. Second, student-level randomization within classrooms creates opportunities for spillovers that may narrow treatment–control contrasts, and our instructor-level assignment reduces this particular risk. Third, the multi-instructor experiment by Sungu et al. (2026) evaluates instructor-facing GenAI rather than student-facing assistance. Our study addresses these gaps by examining course-integrated AI tutor access across disciplines, with particular attention to how its introduction changes students' engagement with existing course activities alongside academic performance.

# 2 Study Context

The current investigation was conducted at the University of Maryland (UMD), a public, four-year university in the Mid-Atlantic United States with approximately 40,000 students (30,000 undergraduate, 10,000 graduate) and 4,500 faculty across 12 colleges representing distinct disciplines. The university's Division of Information Technology provisions a number of custom and licensed generative AI tools, including a VSA chatbot, the utility of which is the focus of the current study. The VSA was developed and made available to select instructors on a preliminary basis beginning in 2024, with the goal of piloting emerging GenAI technology to interested faculty within an instructional course space. After an academic year of limited pilot uses, the current study was motivated by the need to evaluate the VSA's utility and impacts on learning outcomes more broadly prior to considering campus-wide deployment of the tool within the learning management system. This research was approved by UMD's Institutional Review Board (protocol number K-2768).

## 2.1 Virtual Study Assistant

The VSA provides a GPT chatbot (GPT 4.o) with RAG technology for use in course spaces within the institution's ELMS, which hosts the online space for each course at UMD. The VSA's RAG architecture allows it to primarily draw content directly from each course's online materials, including pages, PDFs, text documents, Powerpoint slides, and video transcripts. By drawing on these materials and providing citations for the materials it draws in its responses, the tool is intended to improve the relevance and accuracy of its answers while allowing students to consult the underlying sources. However, it is worth noting that the VSA does use an underlying pre-trained LLM. As such, it is subject to the same limitations of other LLMs (e.g., hallucinations), and while responses are designed to remain grounded in course materials, the VSA may use its pre-trained gen-

erative capabilities to create supporting content that is not sourced from course materials (e.g., generating incorrect answer choices for practice quizzes or generating hints in tutor mode). As the administrator of their VSA, the instructor can control what categories of learning materials and what specific sources are pulled from the course space. When source content is added, removed, or modified, these changes are detected, and the VSA initiates a nightly ingest process where the source information is scanned and processed into a vector database to enable the RAG technology. The technological architecture of the VSA is visualized in Figure A1.

Once activated, the VSA appears as an integrated tool in the sidebar navigation menu of the digital course space. Students can interact with the VSA chatbot by querying it. The VSA restricts its responses to queries that it deems relevant to course content. For example, if a student asks an English literature question to the VSA in a biology class, the VSA would indicate to the student the question is out of scope. The VSA also provides citations if the instructor makes that content public. When a video lecture is cited, a link is provided to the timestamp in the video that has the answer. Figure A2 shows several screenshots of student views of the VSA. Students can provide feedback using a thumbs up or thumbs down response, along with comments that are provided to the instructor. Instructors are able to view students' anonymized prompts and VSA responses in an administrative console, as well as aggregated data related to the frequency and content of student prompts (see Figure A3). Additionally, instructors can freely switch between a tutor mode, which guides students toward an answer, and a direct instruction mode, which provides more explicit explanations.

## 2.2 Experimental Design

The current study took place during the Fall 2025 semester. Instructors of record were recruited beginning prior to the semester and continuing through week 7 of the 16-week semester. Instructors provided their informed consent to participate. We made every

effort to recruit instructors teaching courses from a variety of disciplines and academic levels. Instructors were admitted to the study only if they used a distinct ELMS course page for their section(s), and they indicated that a majority of their course content (e.g., syllabus, readings, videos, and/or assignments) would be posted to that course page, as the VSA requires these materials to function as intended. All students in participating courses were notified and enrolled into the study unless they chose to opt out or were under 18 at study onset. A full timeline of the study rollout is presented in Figure A4.

The study design was a group-randomized trial, so randomization occurred at the instructor level within course blocks. Instructors were first grouped based on teaching the same or similar courses, with randomized assignment to the treatment condition (VSA on) or control condition (VSA off) then occurring within each group, such that courses from each group were represented in both conditions. All students taught by an instructor were assigned to the instructor's condition for that course. Thus, if the instructor was assigned to the treatment condition, all of their students for the participating course had access to the VSA, and if an instructor was assigned to the control condition, none of their students had access to the VSA.

The instructor groupings occurred as follows: (a) If two or more instructors taught different sections of the same course (i.e., an *exact match*), they were put into the same group. Otherwise, groupings were made based on (b) instructors teaching courses in the same level and program of study (e.g., two 300-level English courses), or else, (c) teaching qualitatively similar courses based on an evaluation of course description, format, and student number. For groups of two courses, one instructor was randomly assigned to the treatment condition, and the other to the control condition; for groups of three, two were assigned to treatment, and one to control. In order to maximize sample size, recruitment continued after the semester started. While we attempted to enroll any later enrolled instructors together as groups (according to the criteria described above), there were cases in which a newly enrolled course better fit with an existing group. In these cases, the

instructor was added to the existing group and randomly assigned to the treatment or control condition, regardless of existing assignments within the group.

Table A1 shows the distribution of courses by 13 randomization blocks. Participating courses covered a range of disciplines, including biology, computer science, communication, English, family science, education, information studies, public policy, and mathematics. These courses represent 11 of the university's 12 schools or colleges. Course numbers ranged from the 100 to the 400 level, indicating representation from both lower- and upper-division coursework. Nine blocks comprised sections of the same course (i.e., exact-match), while the remaining four blocks included 12 sections from different courses. The combination of the nine exact-match blocks and the remaining four blocks comprises the study's full sample.

## 2.3 Study Conditions

After an instructor's assignment to the treatment condition, the VSA was integrated as a visible hyperlink within their ELMS course space, and instructors received an email from the research team with instructions for setting up and activating the VSA. If instructors did not activate the VSA, it was unavailable for their students. As our experiment aims to evaluate how VSA access affects student outcomes rather than instructors' uptake, we removed these instructors from our sample ($n$ = 3 instructors). In addition to these initial instructions, instructors in the treatment condition received links to customized how-to articles and video tutorials, as well as optional sample talking points, ELMS announcement language, and syllabus text for sharing information about the VSA with their students. Instructors also received automatic weekly insights emails summarizing student use in their course (e.g., number and general topics of interactions). Treated students received links to existing tutorials for using the tool, including a student-targeted video and guide for using the tool. Beyond suggesting to instructors that they introduce and demonstrate how to use the VSA with their students, the researchers provided no ad-

ditional guidance for how they should incorporate the tool. No intervention occurred in courses in the control condition; instructors were informed of their assignment and were only provided sample talking points and announcement language for discussing the study.

# 3 Data

The study ultimately included 30 instructors and 2,379 students.[2] The first column of Table 1 presents the demographics and characteristics of all participating students and instructors. The average student age was approximately 20 years; 56.5% were female. In terms of reported racial and ethnic characteristics, 42.2% were White, 19.3% Black, 31.3% Asian, 11.4% Hispanic, and 7.7% identified with other racial categories. Students were distributed across undergraduate class years: 22.6% were freshmen, 28.6% sophomores, 26.6% juniors, and 21.3% seniors. First-generation students accounted for 21.9% of the sample, while international students represented 2.9%. Average reported high school and undergraduate GPAs were 4.373 and 3.421, respectively.

Participating instructors had an average age of 49 years, and 42.2% were female. The reported racial and ethnic composition was 50.6% White, 10.2% Black, 7.1% Asian, 10.4% Hispanic, and 26.0% in other racial categories. International instructors accounted for 14.5%. Most instructors were lecturers (72.2%), followed by professors (15.7%) and graduate assistants (10.8%).

## 3.1 Sample Construction

Besides the *full sample*, we also consider an *exact-match sample*. As discussed above, because not every course had multiple instructors, we grouped certain instructors into

[2]The final sample included 34 unique course-sections. Some instructors taught multiple sections of the same course. In those cases, the instructor was in either the treatment or control group for all sections taught.

blocks if we deemed that their courses shared certain attributes (e.g., level of study, offered in the same program, content). The exact-match sample omits these blocks and includes only those cases in which the treated and control instructors taught the exact same courses. As students in these blocks went through largely the same curricula, assessments, and grading scales, the exact-match sample holds course content fixed and helps isolate the impact of VSA from other extraneous course-level factors. The exact-match sample contains 22 instructors and 1,353 students, roughly half the observations of the full sample. We report results for both throughout.

## 3.2 Randomization Balance Checks

The randomization procedure resulted in a balance of course and instructor characteristics (e.g., enrollment, instructor age/race/rank) between the two conditions. Student demographic characteristics (e.g., race, high school GPA) showed slightly more imbalance, likely from student sorting into different instructors. Thus, our preferred models include all demographic variables as controls in the subsequent regression analysis.

Tables 1 and A2 report balance for the two samples. Student characteristics are broadly similar across conditions, although some differences emerge: in the full sample, treated students are more likely to be White and less likely to be Asian, and have slightly lower high school GPAs; in the exact-match sample, they differ in first-generation and international status. This pattern is consistent with student sorting across instructors rather than a failure of randomization, since assignment occurred at the instructor level. Course and instructor characteristics are comparable in the full sample, with the exception of instructor race. Differences on instructor ranks are also sizable, although they are not statistically significant. In the exact-match sample, treated instructors are slightly older and assign fewer homework points; at the same time, they are more balanced regarding race and academic ranks. We therefore include student demographics as controls in all specifications and condition on randomization block throughout.

## 3.3 Outcome variables

*Academic outcomes.* Primary academic outcomes include final grades, which are posted by the instructor on the ELMS course page and use a 0–100 scale, and course letter grade of record converted to the GPA scale. The two measures may diverge both because of their differing scales and because some instructors apply different conversion rubrics. A secondary academic outcome is homework score, which is converted using points earned over points available for assignments in the course space. Note that the researchers did not have visibility into which assignments or exams may have taken place in person, online with a test proctoring service, or online without proctoring, nor did the researchers have insights into instructors' policies around AI use for a given assignment.

*ELMS engagement.* The following measures related to ELMS activity were extracted from the ELMS system: page views (i.e., the total number of times a student viewed a page in ELMS during the semester), active days (i.e., the number of days in the semester that the student had either a page view or participation in ELMS), and total participation, which is the sum of all participated activities, including things such as submitting an assignment, responding to a discussion, or taking a quiz. Total participation depends on the activities required by the instructor in ELMS.

## 3.4 Survey Data

Students and instructors in both conditions received a survey at the end of the semester. The survey assessed students' general use of GenAI across the semester, perceptions of student-instructor interactions (e.g., student questions for instructors; office hours attendance), course and instructor evaluations, and for students in the treatment condition who used the VSA, perceptions of the tool's usefulness, accuracy, and usability. Survey items were predominantly sourced from two prior surveys administered at the university: a campus-wide survey on GenAI in teaching and learning, originally administered

in the 2023-2024 academic year, and an operational survey administered to seven courses that piloted the VSA as part of a usability study in Summer 2025.

The survey was administered via the Qualtrics platform and distributed via invitations sent through the Qualtrics mailer, which automatically records respondents' email addresses, enabling the researchers to link survey responses to other study data sources. The survey occurred simultaneously for all participants, beginning on the second to last week of classes (week 14 of the semester) and remaining open through finals (week 16). Student nonrespondents received up to three reminder emails and two direct reminders which appeared as one-time pop-up messages in ELMS within the participating course page. Instructor nonrespondents received one reminder email per survey. Ten student respondents were randomly selected to win $50, and four instructors were randomly selected to win $500. We achieved response rates of 18% for students and 67% for instructors. We ultimately achieved response rates of 18% for students and 70% for instructors.

### 3.5 Patterns of VSA Usage: Students

We characterize student VSA usage using three measures: the number of interactions, defined as user-generated questions or instructions submitted to the VSA; the number of sessions, each of which may include multiple student prompts and VSA-generated responses; and the number of unique users, defined as distinct students who use the VSA in a given course during the relevant reporting period. A session begins when a student opens the VSA and ends when the browser is closed or after two hours of user inactivity.

Based on Table 2, only 14.63% of treated students ever used the VSA. Figure 1 shows that VSA use was uneven across students, over time, and across courses. Conditional on using the tool at least once, approximately 40% of students had only one session, and most had only a few sessions. The distribution nevertheless has a long right tail, with a small number of students recording more than 20 sessions and some exceeding 40.

Usage also varied over the semester. Total interactions were generally higher in the

later instructional weeks, with pronounced peaks in weeks 10–11 and week 14, and a decline during Thanksgiving week. The daily pattern for one large introductory STEM course (bottom left panel) provides more direct evidence of assessment-related use: the number of unique users increased around exam dates and was typically low between exams. These patterns are consistent with students using the VSA episodically to support exam preparation rather than as a regular study companion, although usage timing alone does not establish the purpose of students' interactions.

Finally, average VSA usage varied substantially across courses. The two heaviest-use courses averaged roughly nine and just over three interactions per student, respectively; both are large, foundational STEM courses taken mainly by first-year students. All other courses averaged fewer than two interactions per student. These differences may reflect variation in the VSA's usefulness for particular course content and designs, the extent of instructor encouragement, and students' needs and preferences.

## 3.6 Patterns of VSA Usage: Instructors

We have limited data on how instructors used insights from the VSA to adjust their instruction. Among survey respondents in the treatment group ($n$ = 10), 30% agreed that they used information from emails about students' VSA use to identify topics for review, and 70% reported monitoring students' questions and answers with the VSA at least somewhat. Although all respondents reported mentioning the VSA to their students, only 40% demonstrated how to use it, and none incorporated it into tasks or assignments. Overall, these findings suggest that instructors engaged with the VSA to some extent but did little to integrate it into their instruction, which may have contributed to the low student engagement described above.

# 4 Methods

We estimate the intention-to-treat (ITT) effects of VSA access. Because randomization occurred at the instructor level within the same or similar courses, all specifications include randomization-block fixed effects. We index observations by student and course-section to accommodate instructors teaching multiple sections. For student $i$ in course-section $c$ within randomization block $b$, we estimate

$$Y_{icb} = \beta \, \text{Treated}_c + \mathbf{X}_i'\boldsymbol{\gamma} + \delta_b + \varepsilon_{icb} \tag{1}$$

where $\text{Treated}_c$ indicates assignment to VSA, $\delta_b$ are randomization-block fixed effects, and $\mathbf{X}_i$ is a vector of pre-treatment student covariates: age, gender, race/ethnicity, class standing, first-generation and international status, high school GPA, and prior undergraduate GPA.[3] By controlling for randomization-block fixed effects, we compare treated and control students only within blocks. Moreover, we control for student covariates to improve precision and to account for the imbalance documented in Tables 1 and A2. We estimate Equation (1) separately for the full and exact-match samples described above, and cluster standard errors at the course-section level.

As our outcome variables are at different scales (e.g., GPA at a 0-4 scale, grades at a 0-100 scale, page counts on a numeric scale, etc), we use and report the raw scales in the analysis first to ease interpretation, and also convert the estimated coefficients using standard deviations among control students to show magnitudes and make the effects comparable across outcomes.

Although the treatment-on-the-treated (TOT) effect is also of policy interest, we focus on ITT estimates because the exclusion restriction required to identify a TOT effect may not hold—that is, assignment to VSA access could affect outcomes not only through VSA use but also by changing students' use of other GenAI tools. As discussed above, students

[3] Missing values of the two GPA measures are set to zero and flagged with missingness indicators.

had access to several university-supported and commercially available GenAI tools during the study period. We do not observe their use of these alternatives and therefore cannot assess whether assignment induced substitution toward or away from them independently of actual VSA use. The availability of alternative tools does not itself violate the exclusion restriction, but the possibility of such an additional pathway complicates the interpretation of TOT estimates. Nonetheless, we explore using the randomization as an instrumental variable for VSA use to probe potential mechanisms of the ITT estimates in Section 7.

# 5 Results

Our primary analysis uses all participating classrooms and accounts for the blocked randomization design. We also report estimates for an exact-match subsample restricted to randomization blocks comprising sections of the same course. This analysis assesses whether the findings persist in more comparable instructional settings, although the restriction also changes the composition of the sample. We use the terms full sample and exact-match sample to differentiate the two analyses.

## 5.1 VSA Uptake

Overall, approximately 15 percent of the students in treated classrooms used the VSA at least once during the study. Figure 4 presents a compiler analysis in the full sample showing which student subgroups were more likely to use the VSA compared to their counterparts. Within-classroom comparisons suggest that VSA uptake varied modestly across student characteristics. Hispanic students were more likely to use the VSA than White students, with an estimated difference of approximately 10 percentage points in both the full and exact-match samples; this association is more precisely estimated in the exact-match sample. Students in the second bottom quartile had higher uptake than stu-

dents in the bottom quartile, but higher high school GPA quartiles were not significantly different in their uptake (note that quartiles here refer to the study population, and not students' GPA quartiles relative to their high school class). In the exact-match sample, there is not any evidence of differential VSA usage by student demographics, including gender, first-generation status, year, high school GPA quartile, or race (Figure A5).

## 5.2 Impact of VSA Access

Figures 2 and 3 compare the distributions of learning and behavioral outcomes across treatment and control students in the full and exact-match samples, respectively. In the full sample, final grades are less concentrated near the upper end of the distribution among treatment students, with more density at lower grades. By contrast, the course GPA and homework-score distributions largely overlap. Differences in academic outcomes are more visible in the exact-match sample, where treatment students have less concentrated distributions near the highest grades, GPAs, and homework scores.

The figures show more consistent differences in ELMS engagement. In both samples, the treatment distributions for page views and active days are shifted toward lower values. The contrast is particularly pronounced for total participation: treatment students' activity is concentrated at relatively low levels, whereas the control distribution has substantially more mass at higher participation levels. The patterns suggest that VSA access was associated with lower engagement with the existing learning management system.

Table 3 reports the corresponding ITT estimates (Equation 1), which capture the impact of being assigned to having VSA access. We present the treatment effect estimates without controls in the first row and with controls in the second row. Note that the full sample results are shown in Panel A, and the exact-match results are shown in Panel B.[4]

[4]Because treatment was assigned at the instructor level, inference relies on a relatively small number of clusters (34), a setting in which cluster-robust standard errors can overstate precision. We therefore also conducted randomization inference, re-randomizing treatment across instructors within randomization block while holding the number of treated instructors per randomization block fixed, and comparing the observed $t$-statistics with their distribution across re-randomizations. Our conclusions are robust to this

In the full sample, there is not significant evidence of adverse treatment effects, likely due to the lack of alignment between courses and grading. In the exact-match sample, there are adverse treatment effects in two of the three achievement variables. Notably, in the model that included control variables, course grades were 4.12 points lower in the treatment condition, which is equivalent to 0.37 standard deviations. Similarly, course letter grades in GPA scale were lower by 0.11 points (0.13 SDs). There is not evidence that VSA access impacted homework scores.

Several aspects of these findings merit emphasis. First, final grades provide a more granular, and arguably more objective, measure of course performance than course GPA and may therefore capture changes that do not translate into differences in GPA, particularly when the latter reflects discrete letter-grade thresholds. Second, the large negative effect size (0.37) in the exact-match sample provides evidence that VSA access adversely affected course performance, which is consistent with several recent research studies that find adverse impacts of AI on student learning in K-12 settings (Bastani et al., 2025; Strömberg et al., 2026), especially for performance on exams or high-stakes testing contexts where AI is not available. Third, the estimated decline in final grade is notable given that only approximately 15% of students who were offered access to the VSA used it, and users averaged approximately four sessions. This contrast raises questions about the mechanisms underlying the results, including whether assignment to the VSA negatively influenced study strategies or engagement with other learning resources.

The effects on ELMS engagement are large and statistically significant in both samples. With controls, VSA access in the exact-match sample reduced page views by 720.96 (0.37 SDs), active days by 7.19 (0.51 SDs), and recorded participations by 26.39 (0.90 SDs). The full sample produced similar treatment effects in page views of 593.08 (0.38 SDs), active days of 9.27 (0.61 SDs), and total participation of 22.51 (0.89 SDs). The large effects on total participation across the two samples are particularly notable. Compared with the other

approach.

two measures that capture the overall ELMS engagement, total participation covers more substantively important learning activities such as homework submission, responding to online questions, and interactions with the instructor and classmates. We further use total homework submission and homework submission on time as additional outcome variables and report the results in Table A3. Across both samples and both outcomes, the coefficients are close to zero and statistically insignificant, suggesting that the reduction in ELMS engagement is more likely to be driven by other substantive learning activities. Overall, the most consistent finding is a substantial reduction in ELMS engagement, accompanied by lower final grades rather than improved academic performance. Note that these estimates do not establish that reduced platform engagement *caused* the academic losses or distinguish reduced study effort from substitution toward VSA or other learning resources, only that reduced ELMS engagement and lower final grades tended to co-occur among treated students.

# 6 Heterogeneity

## 6.1 By Student Characteristics

We first explore heterogeneity by student characteristics. First-generation status is the most consistent dimension of differential treatment effects. In the full sample (Table 4), the effect of VSA access on final grades was 2.87 points more negative (0.28 SDs) for first-generation students than for their peers. The implied treatment effects were $-5.10$ points ($-0.50$ SDs) for first-generation students and $-2.23$ points ($-0.22$ SDs) for other students. This differential was larger in the exact-match sample (Table A4), at $-3.89$ points ($-0.35$ SDs). First-generation students also experienced larger reductions in homework scores: an additional 2.29 points (0.22 SDs) in the full sample and 3.82 points (0.34 SDs) in the exact-match sample. Their additional reductions in ELMS page views were 310.75 views (0.20 SDs) and 440.81 views (0.22 SDs), respectively. These patterns suggest more pro-

nounced adverse academic effects among first-generation students, although statistical precision varies across outcomes and samples.

There is little evidence that treatment effects differed systematically by prior academic achievement. None of the interactions with high school GPA quartiles is statistically significant in either sample, and the estimates show no clear achievement gradient. Gender differences are also generally imprecisely estimated. The exception is an additional reduction of 8.77 recorded ELMS participations (0.30 SDs) among women relative to men in the exact-match sample; this difference is only weakly significant ($p<0.1$) and is not statistically significant in the full sample.

Racial and ethnic heterogeneity is more apparent in platform engagement than in academic performance. Relative to White students, Hispanic students experienced additional reductions of 383.47 page views (0.25 SDs) in the full sample and 526.04 page views (0.27 SDs) in the exact-match sample. Students in the "other race" category experienced additional reductions of 6.35 active days (0.42 SDs) and 6.05 active days (0.43 SDs) in the full sample and exact-match sample, respectively. However, there is no statistically significant racial or ethnic heterogeneity in final grades, course GPA, or homework scores.

Overall, the clearest recurring pattern is greater academic harm among first-generation students, alongside some subgroup differences in platform engagement. Isolated significant findings warrant caution given the number of interactions examined, and reduced ELMS activity need not imply reduced learning effort if VSA substitutes for some platform use.

## 6.2 By Course Discipline

As mentioned above, we group all the participating courses into three buckets of disciplines: STEM, Humanities, and Social Sciences/Education/Policy. Our discipline-specific estimates for the full sample (Table 5) show the clearest evidence of adverse academic effects in humanities courses. VSA access reduced final grades by 2.81 points (0.53 SDs)

and course GPA by 0.12 points (0.27 SDs), both statistically significant at the 1% level. Final-grade effects were also negative in STEM, at −2.21 points (−0.21 SDs), and social sciences/education/policy, at −3.50 points (−0.41 SDs), but these estimates were imprecise. Homework-score effects were small and statistically insignificant across all three groups.

Reductions in ELMS engagement were more consistently evident in STEM and social sciences/education/policy. In STEM courses, VSA access reduced page views by 899.36 (0.47 SDs) and active days by 7.28 (0.46 SDs). In social sciences/education/policy, the corresponding reductions were 490.08 page views (0.48 SDs) and 14.43 active days (0.90 SDs). Despite the larger raw decline in STEM page views, the standardized reductions were nearly identical. The decline in active days, however, was substantially larger in social sciences/education/policy. Neither outcome showed a statistically significant effect in humanities.

Total ELMS participation declined across all three discipline groups: by 24.70 recorded participations (0.97 SDs) in STEM, 26.86 (0.93 SDs) in humanities, and 18.85 (1.05 SDs) in social sciences/education/policy. These represent similarly large standardized reductions, although the humanities estimate is only weakly significant. As all the randomization blocks in Humanities are exact matches, the results are identical for both samples. Results across all the outcomes for the other disciplines are qualitatively similar in the exact-match sample (Table A5).

# 7 Mechanisms

## 7.1 VSA Access and Broader AI Use

The effects reported in Table 3, particularly the substantial reductions in ELMS engagement, are striking given the low uptake of the VSA. As discussed in Section 2.3, students had access to both university-supported and commercial AI tools during the study. As-

signment to VSA access may therefore have affected outcomes not only through use of the VSA itself, but also by changing the broader environment for AI use in treatment classrooms. For example, introducing the VSA may have increased the perceived acceptability of GenAI or encouraged students to use other AI tools. Descriptively, Table 2 shows that in the exact-match sample, lower final grades were not confined to students who used the VSA but were similar to those who did not use the tool. The similarity between treated users and nonusers is consistent with the possibility that the mere presence of the VSA within the ELMS course space changed the broader classroom AI environment, not solely use of the VSA itself.

We first examine this possibility using the student survey. These results require caution for two reasons. First, the overall response rate was low, at 18%. Second, controlling for randomization blocks, treatment-group students were 4.3 percentage points less likely to respond than control-group students ($p < 0.01$). Consequently, comparisons among respondents may reflect differential selection into the survey as well as differences induced by treatment assignment.

Figure 5 presents the unadjusted distributions of responses to a survey item concerning instructor encouragement of GenAI use. Similar shares of treatment- and control-group respondents characterized their instructors as neutral toward GenAI use. However, treatment-group respondents were substantially more likely to report that their instructors encouraged GenAI use or embedded it in the course (41% versus 11%) and less likely to report that GenAI use was prohibited or discouraged.

These descriptive patterns suggest that students in treatment classrooms perceived a more permissive environment for GenAI use. Resources provided to treatment instructors, including written guides, video tutorials, and weekly summaries of students' VSA usage, may have contributed to this environment. However, the survey does not establish whether instructors changed their practices, whether students increased their use of other AI tools, or whether either pathway explains the outcome effects. Moreover, the survey

question concerns GenAI generally rather than the VSA specifically. The small share reporting embedded GenAI use therefore suggests limited perceived curricular integration, but should not be interpreted as a direct measure of students' awareness of VSA access.

We next examine the implications of scaling the assignment effects by treatment-induced VSA uptake. Using randomized assignment to VSA access as an instrumental variable (IV) for using the VSA at least once, we examine the same set of outcomes as discussed in the ITT analysis. A causal interpretation of these estimates requires, among other assumptions, that assignment affects outcomes only through actual VSA use. This exclusion restriction would be violated if assignment also changed instructor practices or students' use of alternative AI tools besides the VSA. We therefore present these estimates as a diagnostic exercise rather than interpret them as established causal effects of VSA use.

Appendix Table A6 presents the results. Assignment induced VSA usage by approximately 15 percent of the students in treated classrooms in the full sample and 13 percent in the exact-match sample. The corresponding first-stage $F$-statistics were 50.5 and 16.5. Although take-up was limited, assignment generated a meaningful first-stage relationship. The IV estimates preserve the direction of the assignment effects but are much larger in magnitude. For example, the estimated effect on final grades is $-20.08$ points in the full sample and $-31.36$ points in the exact-match sample, equivalent to approximately $-1.99$ and $-2.83$ control-group SDs, respectively.

These large magnitudes arise in part mechanically because the IV estimator scales the reduced-form ITT by a relatively small increase in VSA use. Nevertheless, they underscore the strong assumptions required to attribute the full assignment effects to VSA use alone. Combined with the survey evidence of greater perceived instructor encouragement, they motivate considering broader changes in the classroom AI environment as a possible mechanism.

Overall, the evidence is consistent with our conjecture that the ITT effects might operate through broader AI use or instructor practices. Our experiment identifies the effect of

offering the VSA in a setting where other AI tools were already widely available, which is central to interpreting naturalistic evaluations of educational AI. While the ubiquity of alternative tools does not preclude causal evaluation of access interventions, it makes identifying tool-specific mechanisms increasingly difficult.

## 7.2 Engagement with Other Learning Activities

We now explore how access to the VSA might directly reduce student learning through substituting their interactions with other key components of learning—instructors, peers, and learning materials. Results reported in Table 3 suggest that VSA access significantly reduced student engagement with ELMS, especially their active participation ($-0.9$ control SDs). We supplement this analysis by using a few relevant survey items. As the bottom figure in Figure 5 shows, treatment-group respondents were more likely to somewhat or strongly agree with the statement, "I asked the instructor fewer questions about course content" (approximately 57% versus 44%).

Comparing VSA users and those who did not use it within treated classrooms, Figure A6 also points to substitution of using AI for human interaction. The two groups of students reported the largest differences on two items: "I sent the instructor fewer emails or [ELMS] messages" and "I asked the instructor fewer questions about course content." They otherwise reported mostly similar perceptions on other course activities.

## 7.3 Content of Student Chats with the VSA

To investigate how students used the VSA, we developed a framework to classify student requests into five categories: (1) *Direct answer*, encompassing requests for facts, definitions, conceptual explanations, solutions to exercises, or generated text or code; (2) *Practice and test preparation*, including requests for practice problems, quizzes, and study guides, as well as students' answers to and requests for solutions to VSA-generated prac-

tice questions; (3) *Course navigation and logistics*, including questions about the location of course materials, assignment requirements, deadlines, exam coverage, and whether particular topics had been covered; (4) *Feedback on an attempt*, in which students supplied their own reasoning, calculations, writing, or code for evaluation; and (5) *Other or unclear*, encompassing greetings, acknowledgments, experimentation, and interactions whose purpose could not be determined from the available context. Each student turn received a single label corresponding to its primary purpose.

Two researchers from our team independently coded a simple random sample of 300 student requests. We used 100 examples to develop the classification instructions and reserved 200 for held-out evaluation. The two annotators achieve an agreement for 92% of the cases, resulting in a Cohen's Kappa of 0.853. For codes the two annotators did not agree upon, a third team member adjudicated the category. AI coding achieved 86% accuracy across 200 student turns, with a weighted F1 of 0.86 and a macro F1 of 0.73 across five categories. More detailed performance metrics for each coding category are reported in Appendix Table A7. After evaluating the GPT-based classifier against the held-out set, we applied the fixed classifier to the full corpus, using GPT-4.1 mini through the OpenAI API under UMD's enterprise agreement with OpenAI.

Of the 3,460 requests students sent to the VSA, 73.8% were classified as direct-answer requests. Although such requests do not necessarily indicate cognitive offloading, their predominance suggests that students primarily used the VSA to obtain information, explanations, or solutions. Practice and test preparation, which may reflect more self-directed engagement, was the second most common category and accounts for 11.3% of requests. Course navigation and logistics, other or unclear interactions, and feedback on an attempt accounted for 7.6%, 6.6%, and 0.7% of requests, respectively. Requests for feedback on students' own work were thus particularly rare. Overall, while it is difficult to gauge student cognitive engagement with the VSA using our coding categories, the results suggest that students see the VSA as a source of immediate information gathering

rather than as a partner in sustained, interactive learning.

# 8 Discussion

This experimental study investigated the impact of offering a course-integrated GenAI-powered study assistant across multiple disciplines at a large U.S. public university. Over the course of a semester of implementation, we find that access to this tool did not improve academic performance; rather, we find that in some cases, students in the treatment condition had lower final grades. While there was evidence of a statistically significant difference in conditions in full-match sample, the exact-match sample had differences of approximately 0.37 SDs, a difference of about 4 points on a 0-100% scale. The most consistent findings, however, were behavioral: access substantially reduced ELMS page views, active days, and recorded participation. These results extend the evaluation of educational AI beyond performance on assisted tasks to how its introduction changes college students' learning behaviors and engagement with the surrounding learning environment in a naturalistic higher education context.

The behavioral findings merit particular attention. Fewer page views in the learning management system could reflect more efficient information retrieval, which is one goal of the tool's design. However, the decline extended to recorded participation—including assignment submission, discussion responses, and quizzes—which fell by approximately 0.9 SDs in both analytical samples. Meanwhile, as we found no statistically detectable effects on the share of homework submitted or submitted on time, the sizable reduction on total platform participation suggests that students may have lowered their broader engagement with course activities while continuing to meet assignment requirements. The accompanying decline in final grades in the exact-match sample may suggest that the reduced engagement with other learning activities might outweigh potential efficiency gains through the VSA.

Our further analysis of the usage patterns and VSA chats provides descriptive context for these findings. Students' queries to the VSA are primarily composed of direct-answer requests (73.8% of interactions), whereas practice and test preparation (11.3%) and requests for feedback on students' own attempts (7.6%) were much less common. Usage was also episodic and often concentrated around examinations. Together, these patterns suggest that students commonly used the VSA for on-demand information, explanations, and solutions rather than sustained engagement through attempting, receiving feedback, and revising. Survey reports of fewer instructor contacts are also consistent with the substitution in help-seeking.

Our findings complement the small but growing evidence that AI assistance can undermine learning under some conditions (Bastani et al., 2025; Strömberg et al., 2026), while contrasting with gains from deliberately structured AI tutoring (Kestin et al., 2025; Oreopoulos and Low, 2026). One relevant distinction is between technical and instructional integration. Although the VSA retrieved instructor-selected materials and was embedded in the learning management system, instructors did not receive any prescribed approaches for incorporating it into learning activities, and none of the treatment-group instructor survey respondents reported integrating it into assignments. At the same time, while they had the flexibility to choose between direct instruction and tutoring modes, most instructors used the default direct instruction mode. Thus, course grounding through the RAG technology and convenient access may improve the relevance of AI assistance, but it does not ensure productive use.

The interpretation of our results must take the broader AI environment in the university and society into account. Because other AI tools were already available during the study, introducing the VSA could have affected perceived permission to use AI, instructor practices, or students' use of alternative tools. While only approximately 15% of treated students used the VSA, behavior differences between the treatment and control classrooms were substantial. Thus, the intervention likely evaluates an impact of broader

AI use induced by introducing the VSA and not just the isolated effect of individual VSA use.

There are also several limitations of the study. For example, using course grades rather than independently administered assessments makes interpreting a learning effect challenging. In addition, behavior outcomes from interactions with the learning management system only capture a fraction of learning activities, with many others unobservable to the researchers. Despite within-course random assignment, sections may still differ in grading and opportunities for platform participation. That said, it is only through exploration of the real-world deployment of these tools that we can hope to develop a clear-eyed understanding of the myriad factors that might enhance, hinder, or obscure the effect of these tools in the very context for which they were developed. As the first RCT that evaluates the impact of offering AI assistance across many disciplines, the results here highlight a consequential deployment risk: expanding access to AI assistance may change engagement with established learning activities and hinder course performance. Future research should investigate to what extent tool designs and instructional pedagogy can enable AI to be a complement to, rather than a substitute for, existing opportunities for practice, feedback, and human interaction.

# References


Al-Abdullatif, A. M., Al-Dokhny, A. A., and Drwish, A. M. (2023). Implementing the Bashayer chatbot in Saudi higher education: Measuring the influence on students' motivation and learning strategies. *Frontiers in Psychology*, 14:1129070.

Avello-Martínez, R., Gajderowicz, T., and Gómez-Rodríguez, V. G. (2024). Is ChatGPT helpful for graduate students in acquiring knowledge about digital storytelling and reducing their cognitive load? an experiment. *Revista de Educación a Distancia*, 24(78):8.

Barcaui, A. (2025). ChatGPT as a cognitive crutch: Evidence from a randomized controlled trial on knowledge retention. *Social Sciences & Humanities Open*, 12:102287.

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.

Cotton, D. R., Cotton, P. A., and Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in education and teaching international*, 61(2):228–239.

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

Farrokhnia, M., Banihashem, S. K., Noroozi, O., and Wals, A. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in education and teaching international*, 61(3):460–474.

Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences*, 13(7):692.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al. (2023). ChatGPT for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Kestin, G., Miller, K., Klales, A., Milbourne, T., and Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15(1):17458.

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., and Maes, P. (2025). Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv preprint arXiv:2506.08872*, 4:206.

McGrath, C., Farazouli, A., and Cerratto-Pargman, T. (2025). Generative AI chatbots in higher education: A review of an emerging research area. *Higher Education*, 89(6):1533–1549.

Michel-Villarreal, R., Vilalta-Perdomo, E., Salinas-Navarro, D. E., Thierry-Aguilera, R., and Gerardou, F. S. (2023). Challenges and opportunities of Generative AI for higher education as explained by ChatGPT. *Education Sciences*, 13(9):856.

Oreopoulos, P., Liut, M., Sungu, A., and Low, N. (2026). Making AI tutoring productive: Evidence from a mastery-based math practice experiment. Technical report, National Bureau of Economic Research.

Oreopoulos, P. and Low, N. (2026). One click away: AI tutoring with Khanmigo in a two-year school experiment. Technical report, National Bureau of Economic Research.

Purwono, P., Ma'arif, A., Marhoon, H. M., Abougarair, A. J., and Sharkawy, A.-N. (2025). Generative AI as a pedagogical co-pilot in higher education: A review of adaptive learning and teacher support systems. *Journal of Technological Pedagogy and Educational Development*, 2(1):28–42.

Rudolph, J., Tan, S., and Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching*, 6(1).

Stadler, M., Bannert, M., and Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior*, 160:108386.

Strömberg, D., Lei, V., and Wu, Y. (2026). The generative AI learning penalty: Evidence from Chinese secondary education. *Available at SSRN 6868618*.

Sungu, A., Lira, B., and Duckworth, A. (2026). Generative AI can harm teaching. *Available at SSRN*.

Thoeni, A. and Fryer, L. K. (2026). AI chatbots in higher education: Comparing expectations to evidence. *Computers in Human Behavior Reports*, page 101061.

Walczak, K. and Cellary, W. (2023). Challenges for higher education in the era of widespread access to Generative AI. *Economics and Business Review*, 9(2):71–100.

Wang, J. and Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications*, 12(1).

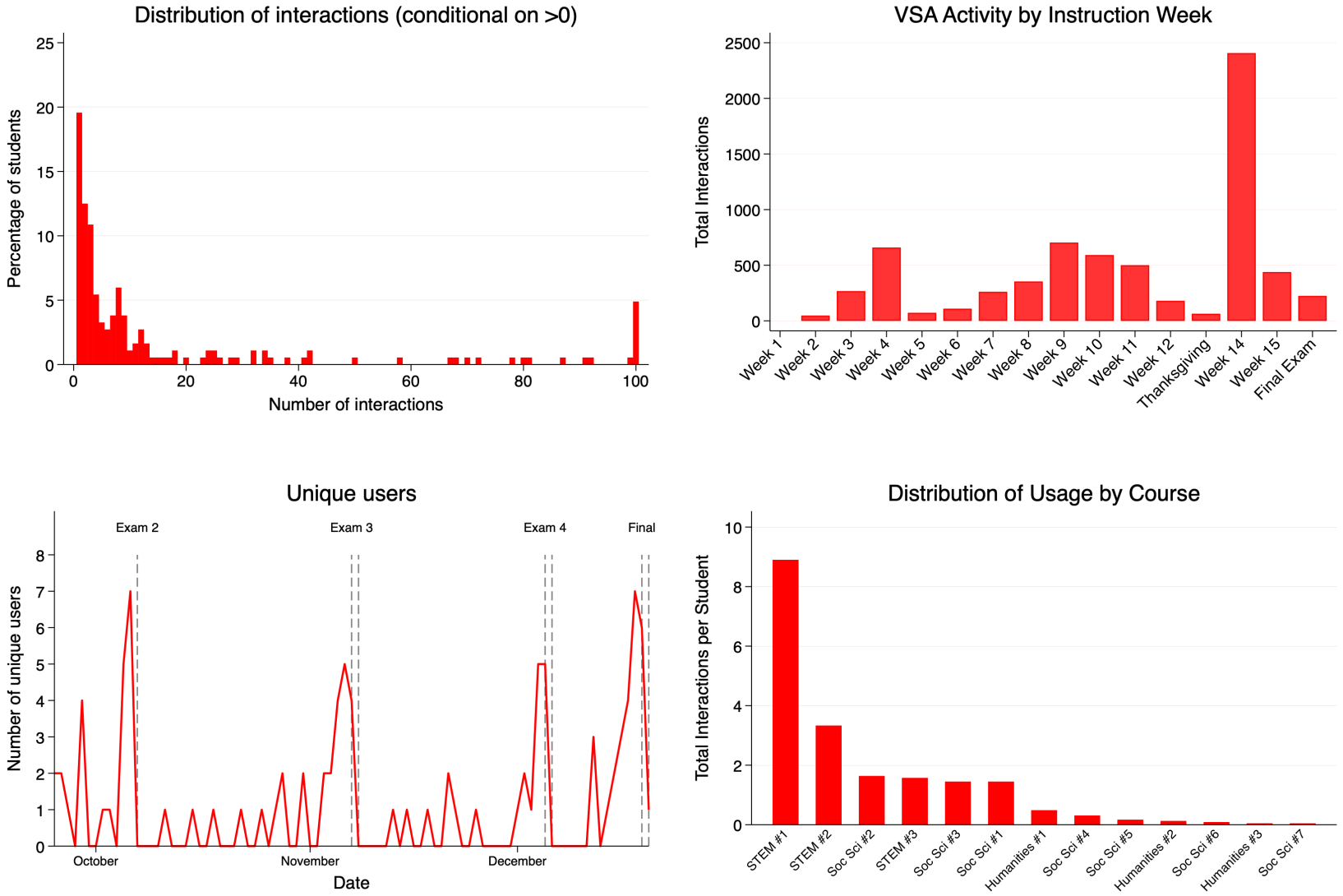
Winkler, R. and Söllner, M. (2018). Unleashing the potential of chatbots in education: A state-of-the-art analysis. In *Academy of management proceedings*, volume 2018, page 15903. Academy of Management Briarcliff Manor, NY 10510.

Wu, T.-T., Lee, H.-Y., Li, P.-H., Huang, C.-N., and Huang, Y.-M. (2024). Promoting self-regulation progress and knowledge construction in blended learning via ChatGPT-based learning aid. *Journal of Educational Computing Research*, 61(8):3–31.

Xu, H., Gan, W., Qi, Z., Wu, J., and Yu, P. S. (2024). Large language models for education: A survey.

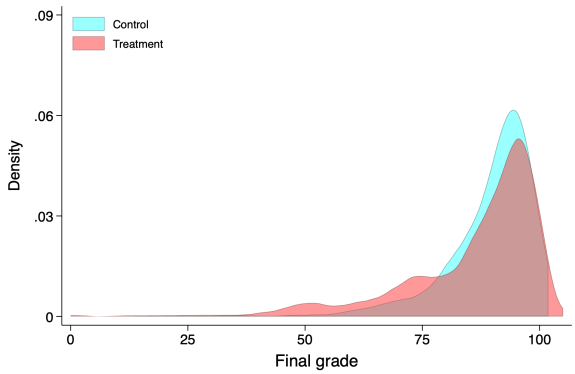
Yu, R., Xu, Z., Ch-Wang, S., and Arum, R. (2024). Whose chatgpt? unveiling real-world educational inequalities introduced by large language models. *arXiv preprint arXiv:2410.22282*.

# Figures and Tables

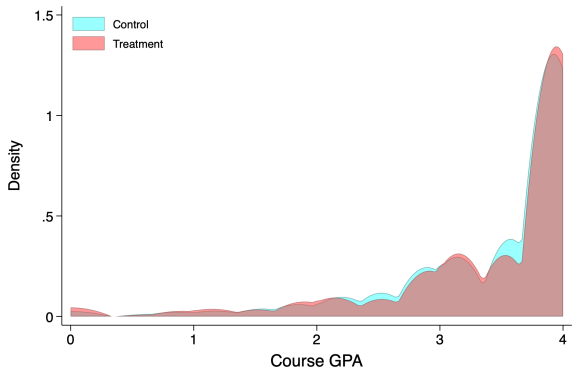



*Notes:* The bottom left panel shows the number of unique VSA users by calendar date for one large introductory STEM course. The pattern regarding usage spikes around exam dates is similar for other participating courses.

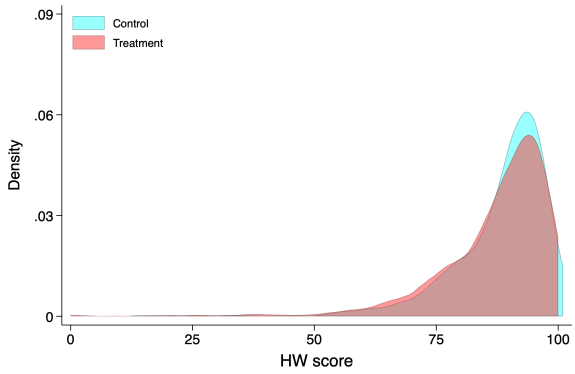
Figure 1: VSA Usage Patterns



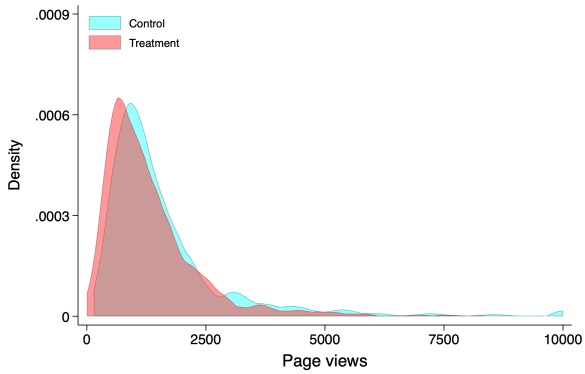

(a) Final Grades



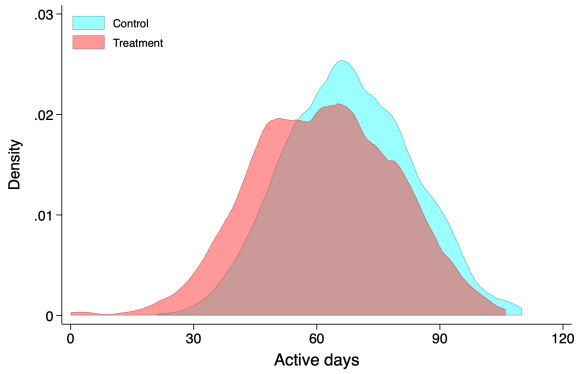

(b) GPA



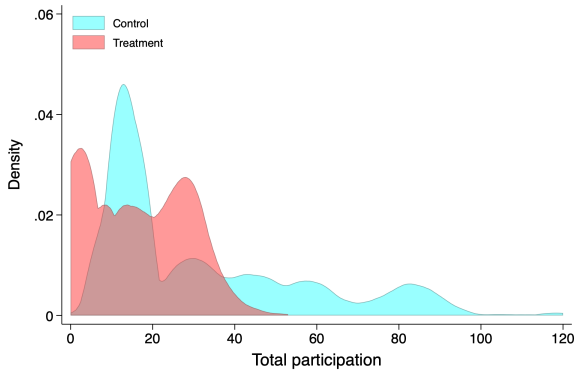

(c) HW Score




(d) Page Views

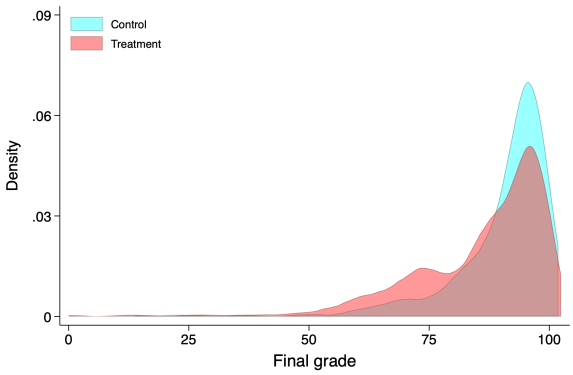



(e) Active Days

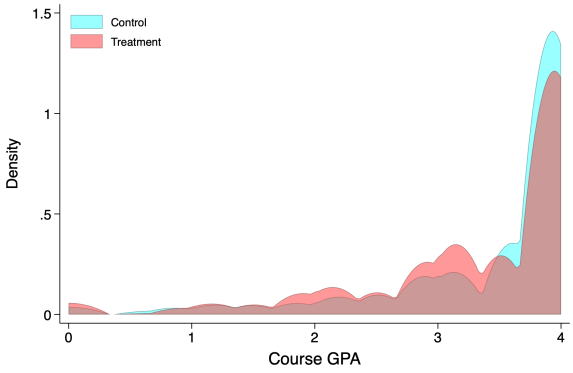



(f) Total Participation

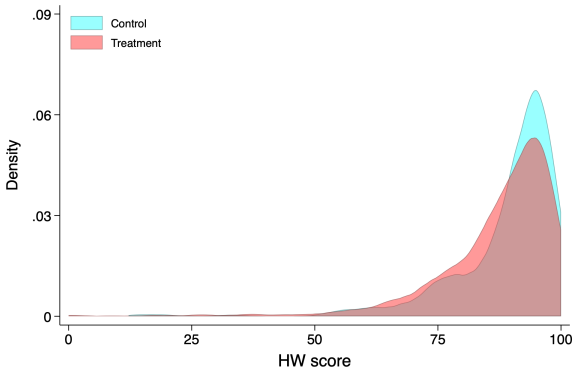
Figure 2: Distribution of Learning and Behavioral Outcomes (Full Sample)



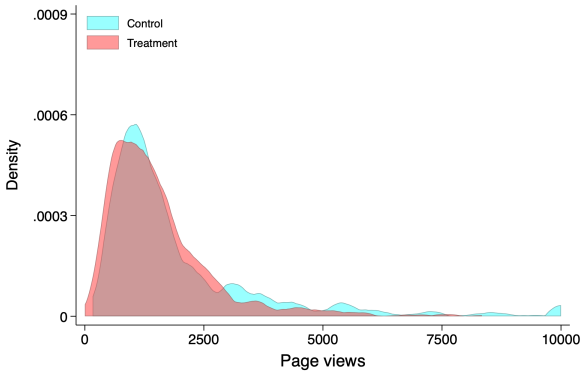

(a) Final Grades



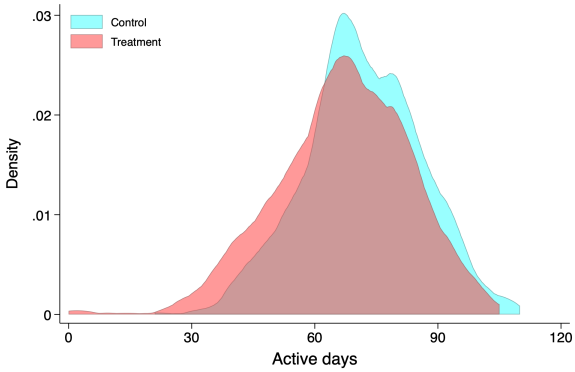

(b) GPA



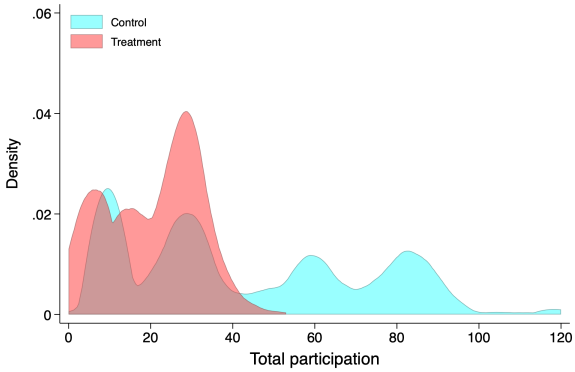

(c) HW Score




(d) Page Views




(e) Active Days




(f) Total Participation

Figure 3: Distribution of Learning and Behavioral Outcomes (Exact-Match Sample)

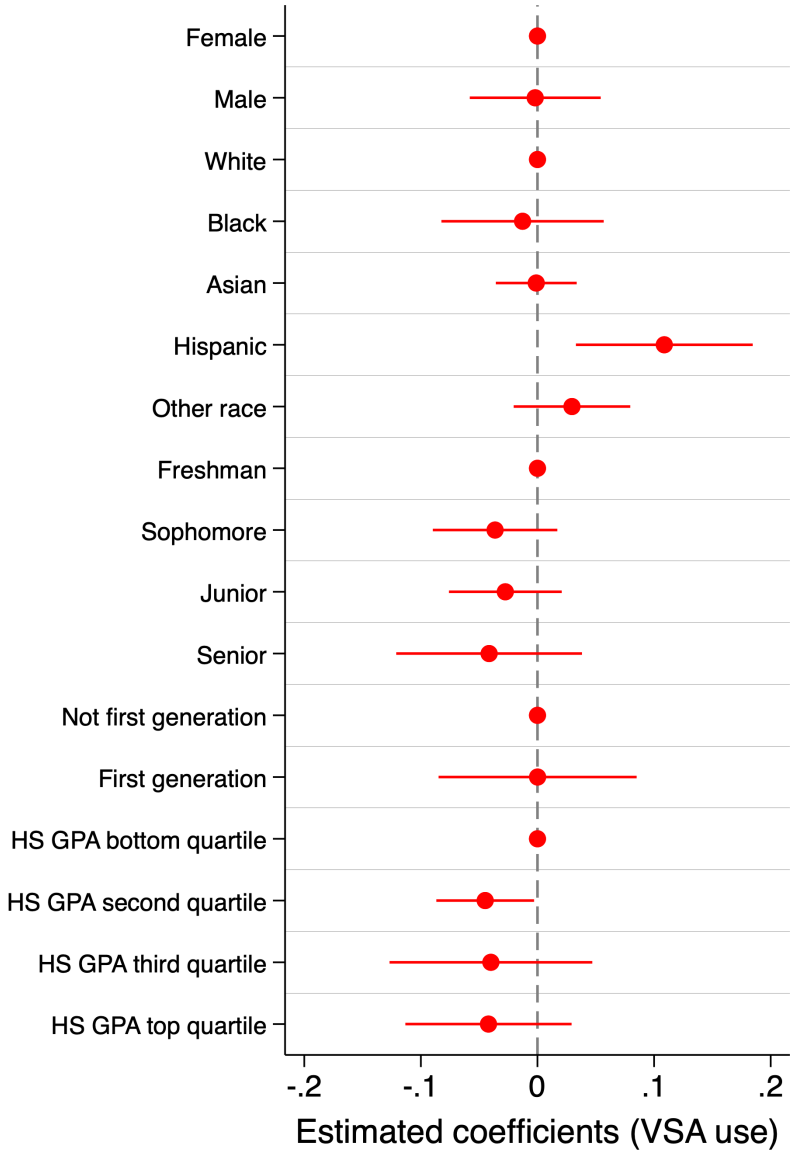



Figure 4: Complier Analysis (Full Sample)

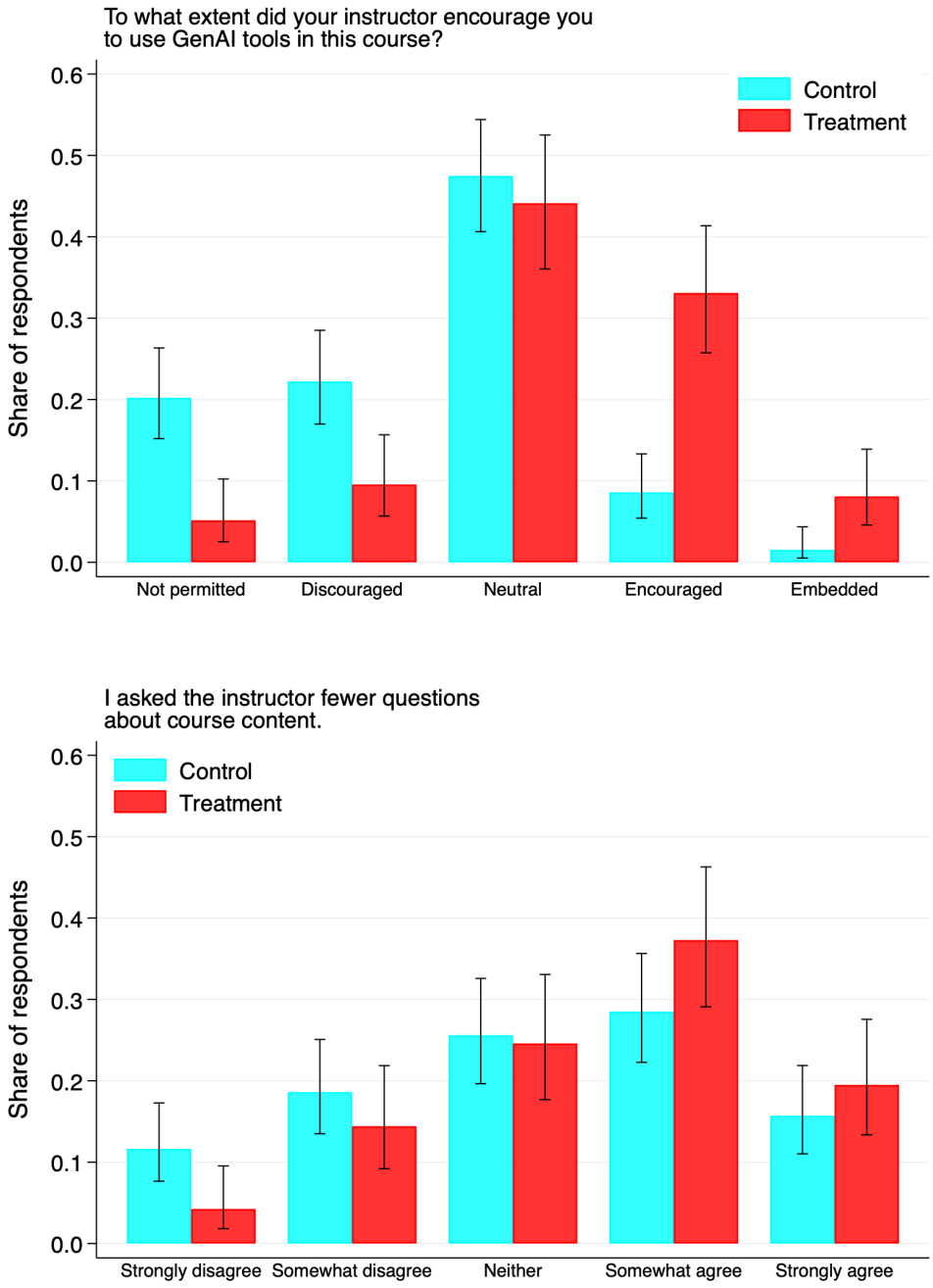



Figure 5: Survey Analysis

Table 1: Sample Balance (Full Sample)

| Variable | Control mean | Treatment mean | Difference (p-value) |
|---|---|---|---|
| **Student Characteristics** | | | |
| Age | 20.185 | 19.744 | 0.352 |
| Female | 0.577 | 0.551 | 0.079 |
| White | 0.370 | 0.478 | 0.001 |
| Black | 0.200 | 0.186 | 0.668 |
| Asian | 0.356 | 0.267 | 0.014 |
| Hispanic | 0.109 | 0.120 | 0.327 |
| Race (Others) | 0.076 | 0.077 | 0.836 |
| Freshman | 0.146 | 0.311 | 0.210 |
| Sophomore | 0.252 | 0.323 | 0.350 |
| Junior | 0.346 | 0.181 | 0.085 |
| Senior | 0.253 | 0.170 | 0.190 |
| First Generation | 0.230 | 0.207 | 0.140 |
| International | 0.025 | 0.034 | 0.084 |
| HS GPA | 4.399 | 4.345 | 0.015 |
| UG GPA | 3.414 | 3.431 | 0.813 |
| Observations | 1,227 | 1,152 | |
| **Course and Instructor Characteristics** | | | |
| Number of Students | 125.807 | 111.498 | 0.947 |
| HW Assigned | 35.374 | 30.835 | 0.129 |
| HW Possible Score | 992.136 | 939.495 | 0.176 |
| Instructor Age | 46.355 | 51.964 | 0.109 |
| Instructor Female | 0.403 | 0.442 | 0.774 |
| Instructor White | 0.672 | 0.330 | 0.075 |
| Instructor Black | 0.068 | 0.138 | 0.138 |
| Instructor Asian | 0.113 | 0.025 | 0.942 |
| Instructor Hispanic | 0.082 | 0.127 | 0.455 |
| Instructor Race (Others) | 0.147 | 0.380 | 0.513 |
| Instructor International | 0.055 | 0.242 | 0.504 |
| Professor | 0.034 | 0.287 | 0.133 |
| Lecturer | 0.800 | 0.639 | 0.269 |
| Graduate Assistant | 0.166 | 0.045 | 0.292 |
| Observations | 17 | 17 | |

*Notes:* This table reports means of pre-treatment characteristics by treatment status for the full sample. Student characteristics are measured at the student level; course and instructor characteristics are measured at the course-section level, with observations reported as the number of sections in each condition. The final column reports the $p$-value from a test of the difference in means between the control and treatment, conditional on randomization-block fixed effects, with standard errors clustered at the course level for student characteristics. High school GPA and undergrad GPA are computed over students with non-missing values. Race/ethnicity categories are not mutually exclusive, since students and instructors may report more than one category.

Table 2: Summary Statistics of Outcome Variables

| | Full sample | | | Exact-match sample | | |
|---|---|---|---|---|---|---|
| | Control | Treatment (used VSA) | Treatment (did not use VSA) | Control | Treatment (used VSA) | Treatment (did not use VSA) |
| *Panel A. Learning outcomes* | | | | | | |
| Final grade | 89.41 | 85.13 | 85.85 | 90.22 | 85.19 | 85.66 |
| | (10.10) | (15.41) | (15.21) | (11.09) | (14.56) | (14.55) |
| GPA | 3.51 | 3.44 | 3.50 | 3.51 | 3.15 | 3.41 |
| | (0.78) | (0.96) | (0.83) | (0.86) | (1.14) | (0.89) |
| HW score | 88.24 | 86.56 | 87.09 | 89.03 | 86.38 | 87.00 |
| | (10.18) | (11.99) | (12.35) | (11.36) | (13.51) | (12.87) |
| *Panel B. Learning behaviors* | | | | | | |
| ELMS page views | 1654.56 | 1338.92 | 1326.18 | 2004.33 | 1696.37 | 1525.54 |
| | (1559.40) | (1190.17) | (1012.36) | (1969.14) | (1380.98) | (1102.98) |
| ELMS active days | 67.79 | 63.94 | 61.23 | 71.67 | 71.56 | 66.26 |
| | (15.29) | (17.25) | (17.39) | (14.16) | (14.84) | (16.40) |
| ELMS total participation | 31.99 | 10.73 | 16.64 | 43.40 | 15.88 | 21.39 |
| | (25.36) | (11.30) | (12.27) | (29.25) | (11.53) | (10.99) |
| HW submitted | 0.91 | 0.91 | 0.89 | 0.92 | 0.92 | 0.91 |
| | (0.11) | (0.12) | (0.13) | (0.11) | (0.12) | (0.11) |
| HW submitted on-time | 0.85 | 0.88 | 0.84 | 0.85 | 0.89 | 0.85 |
| | (0.16) | (0.13) | (0.16) | (0.16) | (0.14) | (0.16) |
| Observations | 1,227 | 172 | 980 | 612 | 97 | 644 |

*Notes:* This table reports means of the outcomes studied in the paper, with standard deviations in parentheses. Treated students are split by whether they used the VSA at least once during the semester. Final grade and HW score are on a 0–100 scale; GPA is on a 0–4 scale; HW submitted and HW on-time are shares of assigned homework; page views, active days, and total participation are counts over the semester.

Table 3: The Effects of VSA Access

| | (1) Final grades | (2) Course GPA | (3) HW score | (4) Page views | (5) Active days | (6) Total participation |
|---|---|---|---|---|---|---|
| *Panel A. Full sample* | | | | | | |
| *Without controls* | | | | | | |
| Treated | -2.68* | 0.05 | -0.74 | -566.08*** | -9.43*** | -21.37*** |
| | (1.54) | (0.07) | (0.72) | (153.92) | (1.57) | (4.07) |
| *With controls* | | | | | | |
| Treated | -2.88* | 0.02 | -0.92 | -593.08*** | -9.27*** | -22.51*** |
| | (1.60) | (0.06) | (0.79) | (159.97) | (1.56) | (4.18) |
| Control SD | 10.10 | 0.78 | 10.18 | 1559.40 | 15.29 | 25.36 |
| Observations | 2,379 | 2,362 | 2,379 | 2,379 | 2,379 | 2,379 |
| *Panel B. Exact-match sample* | | | | | | |
| *Without controls* | | | | | | |
| Treated | -3.55* | -0.07** | -1.12 | -693.54** | -7.10*** | -26.09*** |
| | (1.90) | (0.03) | (0.67) | (251.19) | (1.54) | (6.25) |
| *With controls* | | | | | | |
| Treated | -4.12** | -0.11*** | -1.54* | -720.96*** | -7.19*** | -26.39*** |
| | (1.80) | (0.02) | (0.76) | (230.17) | (1.69) | (6.02) |
| Control SD | 11.09 | 0.86 | 11.36 | 1969.14 | 14.16 | 29.25 |
| Observations | 1,353 | 1,343 | 1,353 | 1,353 | 1,353 | 1,353 |

*Notes:* This table reports the effects of assignment to VSA access on the outcome indicated in each column. Each cell is from a separate regression. All specifications include randomization-block fixed effects; the *With controls* specifications additionally include student covariates: age, gender, race/ethnicity, class standing, first-generation and international status, high school GPA, and prior undergraduate GPA, with missing GPA values set to zero and flagged with missingness indicators. Standard errors clustered at the course-section level, the unit of randomization, are in parentheses. $^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

Table 4: Heterogeneous Treatment Effects (Full Sample)

| | (1) Final grades | (2) Course GPA | (3) HW score | (4) Page views | (5) Active days | (6) Total participation |
|---|---|---|---|---|---|---|
| *High school GPA* | | | | | | |
| Treated (Bottom quartile) | -3.44* | -0.01 | -1.24 | -616.36** | -8.59*** | -21.10*** |
| | (2.03) | (0.09) | (1.43) | (256.62) | (2.06) | (3.95) |
| Treated × Second quartile | -0.25 | 0.02 | -0.08 | 36.81 | -0.40 | -1.23 |
| | (1.54) | (0.11) | (1.61) | (190.25) | (1.83) | (1.20) |
| Treated × Third quartile | 0.76 | 0.04 | 0.65 | -86.42 | 0.05 | -3.11 |
| | (1.13) | (0.09) | (1.12) | (114.31) | (1.46) | (2.20) |
| Treated × Top quartile | 0.26 | 0.02 | 0.05 | 127.57 | 0.26 | -3.26 |
| | (2.02) | (0.15) | (1.87) | (191.11) | (1.84) | (2.14) |
| Observations | 1,967 | 1,952 | 1,967 | 1,967 | 1,967 | 1,967 |
| *Gender* | | | | | | |
| Treated (Male) | -2.69* | -0.01 | -0.90 | -446.18*** | -9.84*** | -20.76*** |
| | (1.32) | (0.06) | (0.95) | (99.93) | (1.74) | (3.98) |
| Treated × Female | -0.35 | 0.07 | -0.04 | -264.13 | 1.03 | -3.15 |
| | (1.25) | (0.06) | (1.03) | (170.36) | (1.51) | (3.09) |
| Observations | 2,379 | 2,362 | 2,379 | 2,379 | 2,379 | 2,379 |
| *Race/ethnicity* | | | | | | |
| Treated (White) | -2.92 | 0.04 | -0.71 | -521.91*** | -7.38*** | -23.37*** |
| | (1.96) | (0.08) | (0.86) | (142.90) | (1.37) | (4.72) |
| Treated × Black | 0.40 | 0.05 | -0.17 | -125.17 | -2.99 | 1.55 |
| | (1.46) | (0.10) | (1.46) | (137.43) | (1.87) | (2.08) |
| Treated × Asian | -0.48 | -0.08 | -0.27 | -92.62 | -2.94 | 1.67 |
| | (1.10) | (0.07) | (1.07) | (94.52) | (1.74) | (1.58) |
| Treated × Hispanic | -0.73 | 0.02 | -1.73 | -383.47** | -2.04 | 2.09 |
| | (3.13) | (0.20) | (2.78) | (177.99) | (2.52) | (2.66) |
| Treated × Other race | 1.50 | -0.06 | -0.33 | -44.12 | -6.35** | 0.34 |
| | (1.42) | (0.10) | (1.09) | (160.71) | (2.54) | (2.63) |
| Observations | 2,379 | 2,362 | 2,379 | 2,379 | 2,379 | 2,379 |
| *First-generation status* | | | | | | |
| Treated (Not first gen.) | -2.23 | 0.04 | -0.41 | -523.13*** | -8.76*** | -22.34*** |
| | (1.67) | (0.07) | (0.83) | (144.59) | (1.58) | (4.31) |
| Treated × First generation | -2.87** | -0.07 | -2.29* | -310.75** | -2.27 | -0.76 |
| | (1.19) | (0.08) | (1.16) | (127.35) | (1.58) | (1.89) |
| Observations | 2,379 | 2,362 | 2,379 | 2,379 | 2,379 | 2,379 |

*Notes:* Each block reports a separate regression interacting treatment with the indicated characteristic. The first row of each block is the treatment effect for the reference group, "Treated ×" rows report differences in treatment effects relative to the reference group. All specifications include randomization-block fixed effects and student covariates. Standard errors clustered at the course-section level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Treatment Effects by Discipline (Full Sample)

| | (1) Final grades | (2) Course GPA | (3) HW score | (4) Page views | (5) Active days | (6) Total participation |
|---|---|---|---|---|---|---|
| *Panel A. STEM* | | | | | | |
| Treated | -2.21 | 0.10 | -0.97 | -899.36** | -7.28*** | -24.70*** |
| | (2.94) | (0.12) | (0.90) | (303.36) | (1.11) | (6.61) |
| Control SD | 10.74 | 0.88 | 10.79 | 1905.64 | 15.67 | 25.53 |
| Observations | 1,253 | 1,239 | 1,253 | 1,253 | 1,253 | 1,253 |
| *Panel B. Humanities* | | | | | | |
| Treated | -2.81** | -0.12** | -0.72 | -149.44 | -2.55 | -26.86* |
| | (0.96) | (0.05) | (1.14) | (98.11) | (1.44) | (13.04) |
| Control SD | 5.35 | 0.44 | 7.06 | 933.69 | 11.20 | 28.75 |
| Observations | 401 | 399 | 401 | 401 | 401 | 401 |
| *Panel C. Social Sciences/Education/Policy* | | | | | | |
| Treated | -3.50 | 0.01 | -0.12 | -490.08*** | -14.43*** | -18.85*** |
| | (2.89) | (0.07) | (1.79) | (106.95) | (2.94) | (4.80) |
| Control SD | 8.57 | 0.60 | 9.28 | 1014.98 | 16.03 | 17.93 |
| Observations | 725 | 724 | 725 | 725 | 725 | 725 |

*Notes:* This table reports the effects of assignment to VSA access on the outcome indicated in each column, estimated separately for each discipline group. Each cell is from a separate regression on the subsample of students in the indicated discipline. Courses are assigned to discipline groups as listed in Table A1. All specifications include randomization-block fixed effects and student covariates. Control SD denotes the standard deviation of the outcome among control students in the corresponding discipline group. Standard errors clustered at the course-section level are in parentheses. $^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$.

# Appendix

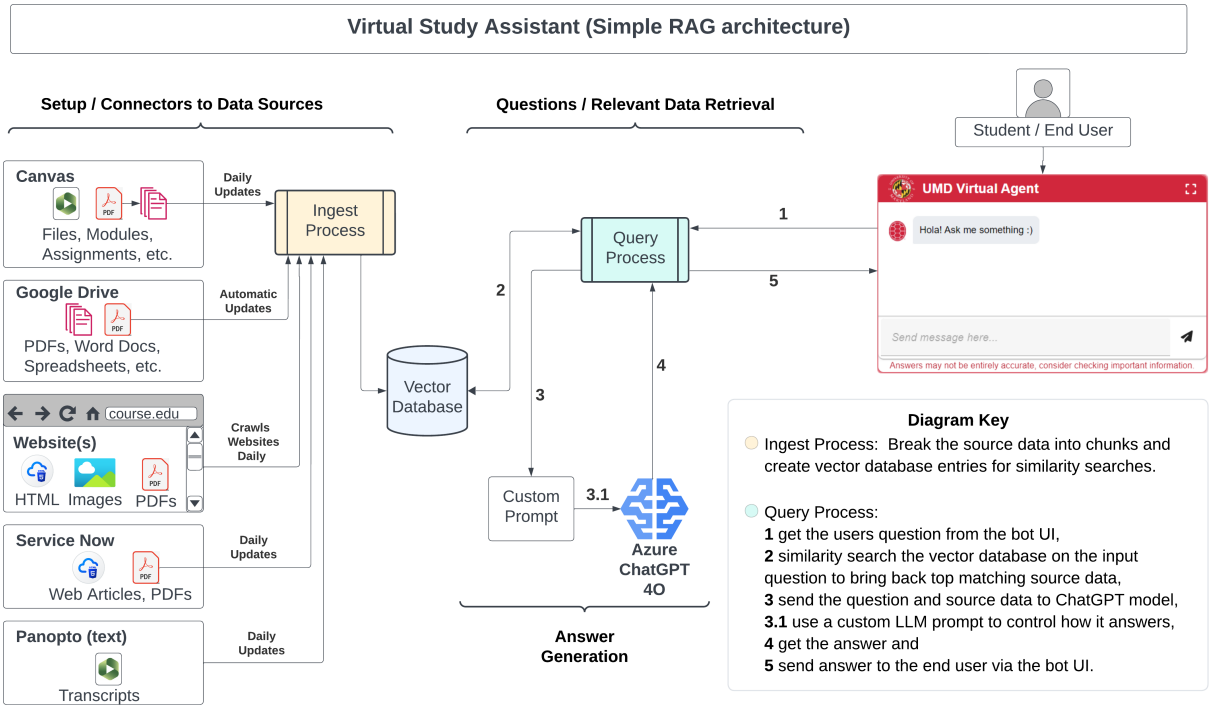



Figure A1: Architecture of the Virtual Study Assistant

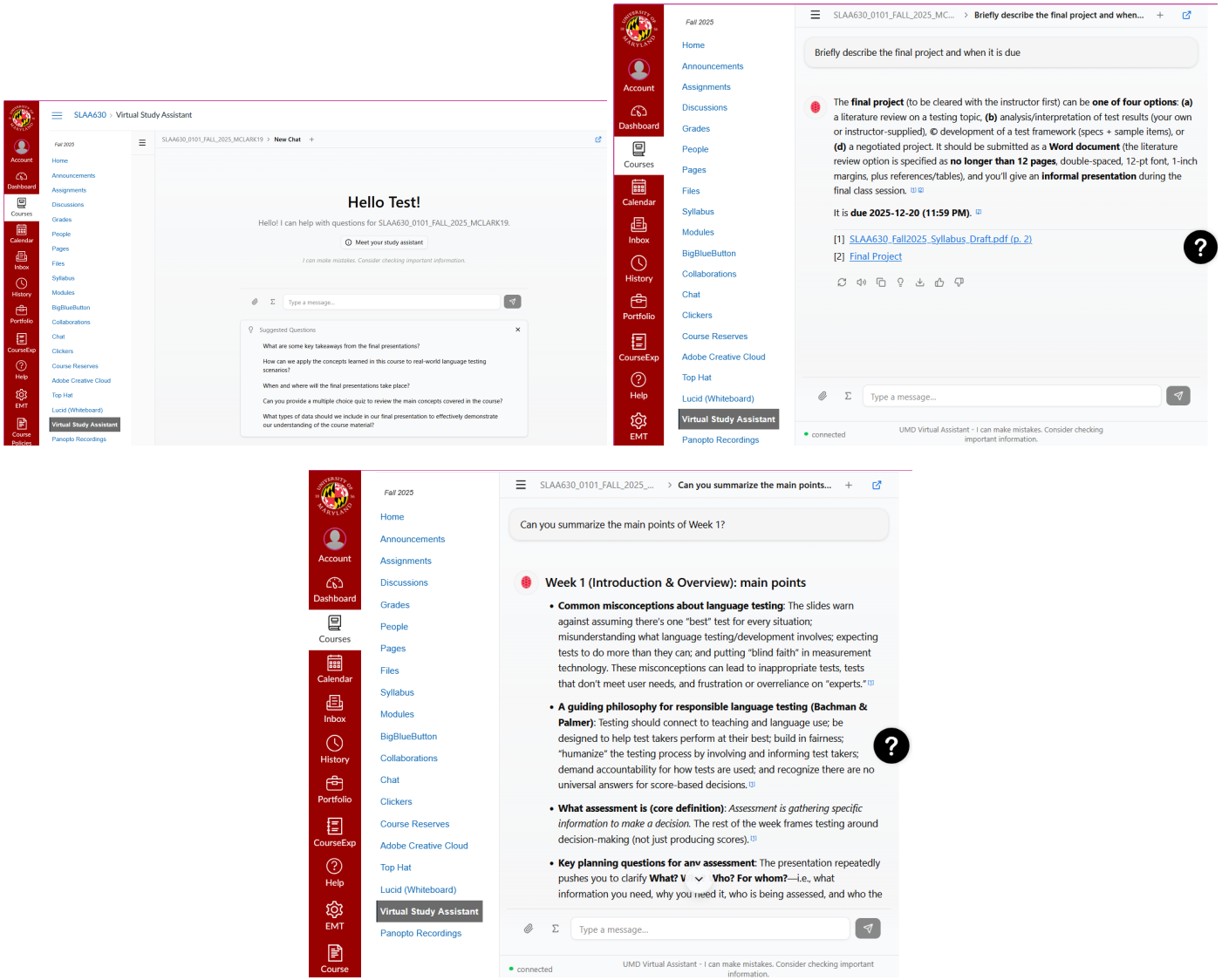



Figure A2: VSA Student Views

Here is a summary of the new insights revealed by your virtual agent from last week, covering student interactions with the **SLAA630** course between **Nov 16, 2025 and Nov 23, 2025**.

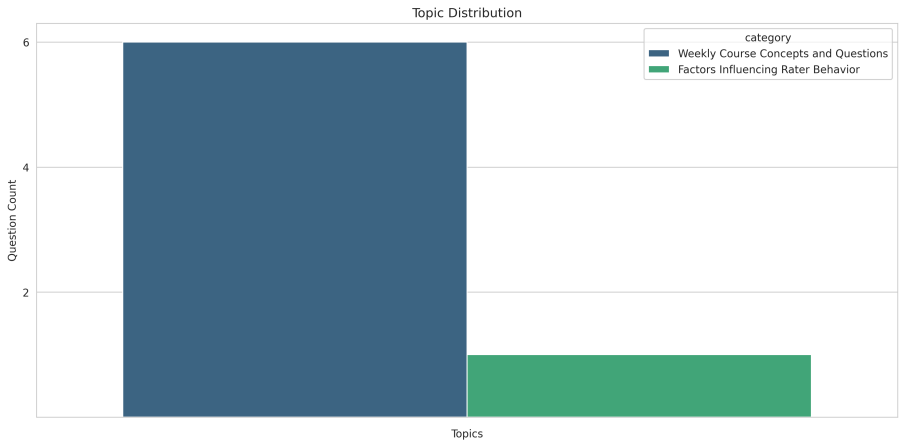
### 1. Top AI-Identified Topics: ✨

Students asked a variety of questions via the virtual agent last week. The main topics generated by AI from these interactions are below.

| Topic | Percentage |
|---|---|
| Factors Influencing Rater Behavior | 14.29% |
| Weekly Course Concepts and Questions | 85.71% |

### 2. Topic Distribution

A visual representation of the main topics generated by AI from these interactions is shown below.




### 3. Questions Receiving Thumbs Down:

No questions received a thumbs-down response this week. 👍🎉

### 4. Negative Sentiment Questions: ☹

This week, no questions were identified as having negative sentiments.😊✨

### 5. Student Interaction Distribution

Below is a visual representation of the number of questions each (anonymized) student has asked the Virtual Agent this past week.

| Number of Questions | Number of Students |
|---|---|
| 0 | 6 |

This report aims to provide insight into the performance of the Virtual Agent for the **SLAA630** course and identify areas that may need improvement. For more detailed insights and analytics, please visit the virtual agent analytics page here.

Figure A3: VSA Instructor Insights

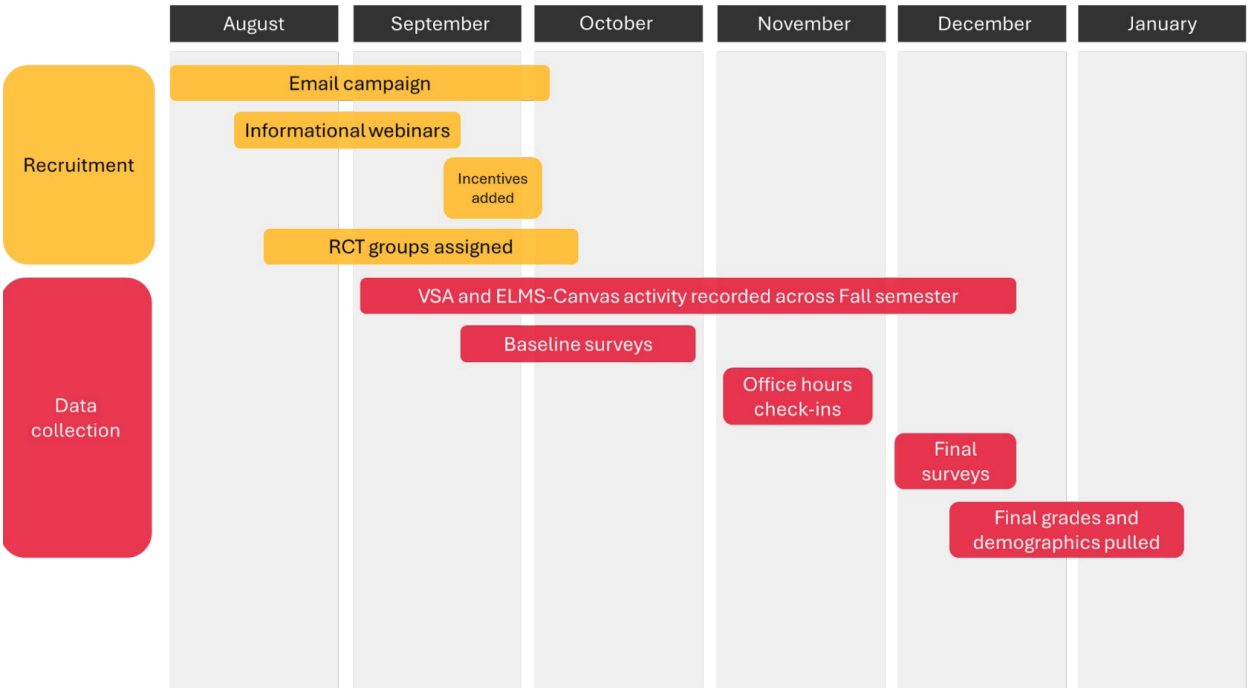



Figure A4: Timeline of Intervention

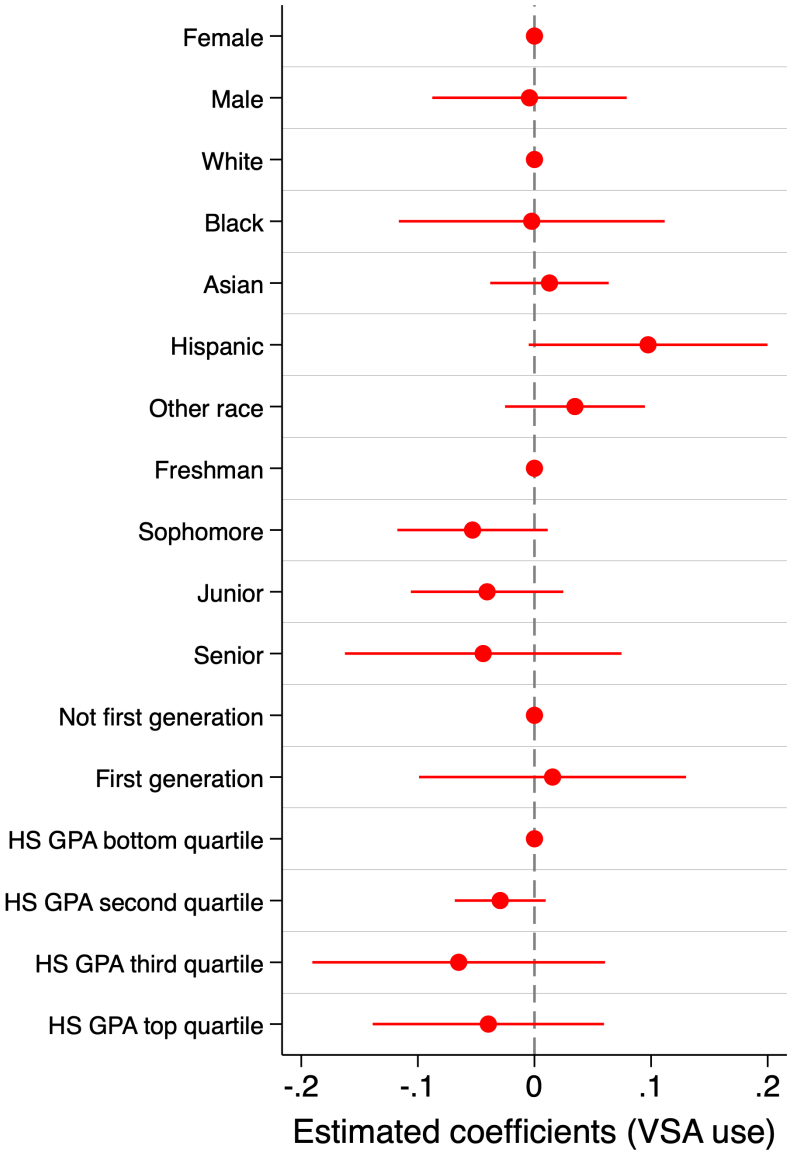



Figure A5: Complier Analysis (Exact-Match Sample)

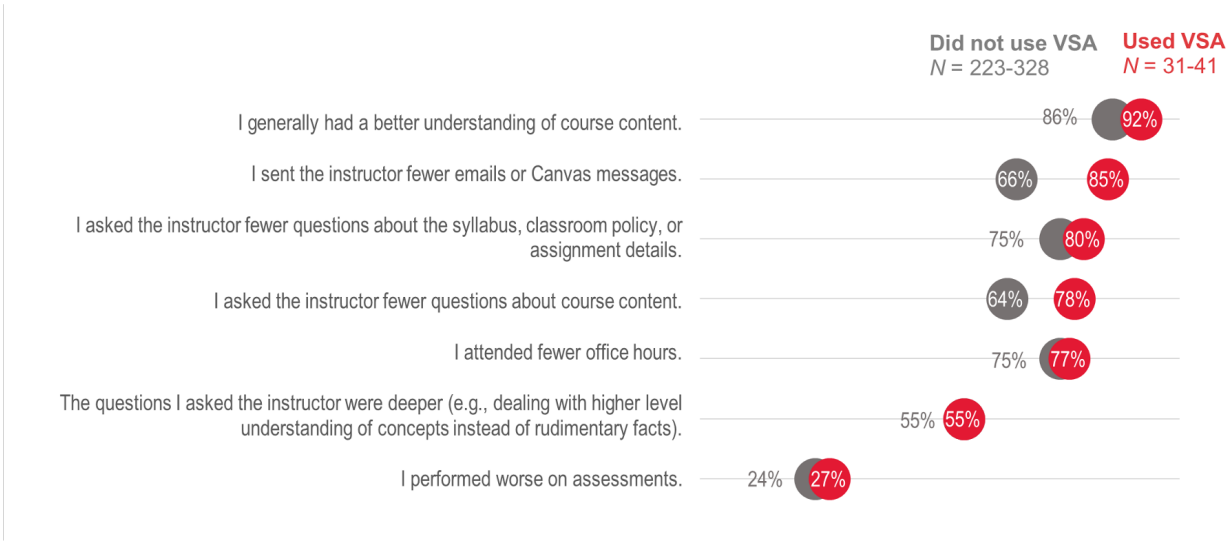



Figure A6: Learning Activites between VSA Users vs. Non-Users

Table A1: Randomization Blocks and Participating Courses

| Block | Courses | Num. of Classrooms | Num. of students | Exact match | Field of study |
|---|---|---|---|---|---|
| A | STEM #3 | 3 | 383 | Yes | STEM |
| B | STEM #2 | 2 | 227 | Yes | STEM |
| C | Humanities #2 | 3 | 68 | Yes | Humanities |
| D | Humanities #2 | 2 | 52 | Yes | Humanities |
| E | Humanities #3 | 3 | 105 | Yes | Humanities |
| F | Soc Sci #7 | 2 | 87 | Yes | Social Sci./Educ./Policy |
| G | Humanities #1 | 3 | 176 | Yes | Humanities |
| H | Soc Sci #6 | 2 | 100 | Yes | Social Sci./Educ./Policy |
| J | Soc Sci #5 | 2 | 155 | Yes | Social Sci./Educ./Policy |
| K | Soc Sci #1, #2, #8, #9, #12 | 5 | 134 | No | Social Sci./Educ./Policy |
| L | Soc Sci #4, #10 | 2 | 190 | No | Social Sci./Educ./Policy |
| O | Soc Sci #3, #11 | 2 | 59 | No | Social Sci./Educ./Policy |
| P | STEM #1, STEM #4 ($\times$2) | 3 | 643 | No | STEM |

*Notes:* Each block is a randomization group within which instructors were randomly assigned to treatment or control. Courses are labeled by field of study. "Soc Sci" denotes courses in the social sciences, education, and public policy. Classrooms is the number of distinct course sections in the block. Number of students is total number of students across those sections. Exact match indicates blocks in which all sections are of the same course, which constitute the exact-match sample.

Table A2: Sample Balance (Exact-Match Sample)

| Variable | Control mean | Treatment mean | Difference (p-value) |
|---|---|---|---|
| **Student Characteristics** | | | |
| Age | 19.596 | 19.528 | 0.211 |
| Female | 0.507 | 0.538 | 0.569 |
| White | 0.374 | 0.453 | 0.050 |
| Black | 0.208 | 0.163 | 0.172 |
| Asian | 0.322 | 0.290 | 0.141 |
| Hispanic | 0.119 | 0.123 | 0.489 |
| Race (Others) | 0.092 | 0.092 | 0.941 |
| Freshman | 0.279 | 0.312 | 0.802 |
| Sophomore | 0.397 | 0.379 | 0.870 |
| Junior | 0.216 | 0.169 | 0.571 |
| Senior | 0.108 | 0.127 | 0.674 |
| First Generation | 0.229 | 0.185 | 0.000 |
| International | 0.028 | 0.045 | 0.008 |
| HS GPA | 4.363 | 4.380 | 0.244 |
| UG GPA | 3.327 | 3.475 | 0.206 |
| Observations | 612 | 741 | |
| **Course and Instructor Characteristics** | | | |
| Number of Students | 88.232 | 89.726 | 0.699 |
| HW Assigned | 42.876 | 34.120 | 0.065 |
| HW Possible Score | 1385.085 | 926.823 | 0.085 |
| Instructor Age | 39.618 | 54.699 | 0.043 |
| Instructor Female | 0.655 | 0.596 | 0.906 |
| Instructor White | 0.614 | 0.513 | 0.268 |
| Instructor Black | 0.103 | 0.050 | 0.388 |
| Instructor Asian | 0.227 | 0.000 | 0.281 |
| Instructor Hispanic | 0.165 | 0.197 | 0.472 |
| Instructor Race (Others) | 0.056 | 0.240 | 0.193 |
| Instructor International | 0.056 | 0.070 | 0.458 |
| Professor | 0.069 | 0.050 | 1.000 |
| Lecturer | 0.704 | 0.880 | 0.863 |
| Graduate Assistant | 0.227 | 0.070 | 0.805 |
| Observations | 10 | 12 | |

*Notes:* This table replicates Table 1 using the exact-match sample, which restricts to randomization blocks in which treated and control instructors taught sections of the same course.

Table A3: The Effects of VSA Access on Homework Submission

| | (1) HW submitted | (2) HW submitted on time |
|---|---|---|
| *Panel A. Full sample* | | |
| Treated | -0.010 | -0.001 |
| | (0.015) | (0.014) |
| Control SD | 0.110 | 0.153 |
| Observations | 2,379 | 2,379 |
| *Panel B. Exact-match sample* | | |
| Treated | -0.008 | -0.004 |
| | (0.011) | (0.015) |
| Control SD | 0.105 | 0.162 |
| Observations | 1,353 | 1,353 |

*Notes:* This table reports the effects of assignment to VSA access on the share of assigned homework submitted and the share submitted on time. Each cell is from a separate regression. All specifications include randomization-block fixed effects and student covariates. Control SD denotes the standard deviation of the outcome among control students in the corresponding sample. Standard errors clustered at the course-section level are in parentheses. $^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$.

Table A4: Heterogeneous Treatment Effects (Exact-Match Sample)

| | (1) Final grades | (2) Course GPA | (3) HW score | (4) Page views | (5) Active days | (6) Total participation |
|---|---|---|---|---|---|---|
| *High school GPA* | | | | | | |
| Treated (Bottom quartile) | -4.77* | -0.12 | -2.08 | -937.84** | -6.92*** | -25.08*** |
| | (2.37) | (0.11) | (1.46) | (389.42) | (2.19) | (5.58) |
| Treated × Second quartile | -1.38 | -0.10 | -0.65 | 352.81 | 1.69 | -0.39 |
| | (2.41) | (0.16) | (2.55) | (291.46) | (2.37) | (1.99) |
| Treated × Third quartile | 0.13 | 0.01 | 1.02 | 37.13 | -0.20 | -4.94 |
| | (1.03) | (0.11) | (1.13) | (166.54) | (1.76) | (3.63) |
| Treated × Top quartile | 0.51 | -0.03 | 0.51 | 457.19 | 1.25 | -4.12 |
| | (2.52) | (0.21) | (2.29) | (339.45) | (1.79) | (3.07) |
| Observations | 1,119 | 1,111 | 1,119 | 1,119 | 1,119 | 1,119 |
| *Gender* | | | | | | |
| Treated (Male) | -3.08** | -0.10** | -0.99 | -519.02*** | -7.50*** | -21.74*** |
| | (1.24) | (0.04) | (0.77) | (131.55) | (1.77) | (5.91) |
| Treated × Female | -1.96 | -0.02 | -1.04 | -381.54 | 0.58 | -8.77* |
| | (1.77) | (0.06) | (1.30) | (265.91) | (1.27) | (4.41) |
| Observations | 1,353 | 1,343 | 1,353 | 1,353 | 1,353 | 1,353 |
| *Race/ethnicity* | | | | | | |
| Treated (White) | -3.77* | -0.10 | -0.83 | -649.25*** | -5.93*** | -29.58*** |
| | (2.07) | (0.06) | (1.14) | (186.19) | (1.71) | (6.38) |
| Treated × Black | -1.55 | 0.00 | -2.19 | -194.02 | -2.45 | 5.28* |
| | (1.46) | (0.13) | (1.79) | (238.35) | (2.73) | (2.90) |
| Treated × Asian | -0.89 | -0.06 | -0.66 | -32.18 | -0.43 | 5.77** |
| | (1.25) | (0.10) | (1.34) | (162.24) | (1.45) | (2.24) |
| Treated × Hispanic | 0.12 | 0.17 | -1.89 | -526.04* | -2.62 | 5.68 |
| | (4.15) | (0.26) | (3.83) | (260.67) | (3.52) | (4.53) |
| Treated × Other race | 1.54 | 0.02 | -0.42 | 50.57 | -6.05* | 4.69 |
| | (1.86) | (0.14) | (1.57) | (225.71) | (3.07) | (3.92) |
| Observations | 1,353 | 1,343 | 1,353 | 1,353 | 1,353 | 1,353 |
| *First-generation status* | | | | | | |
| Treated (Not first gen.) | -3.32* | -0.09** | -0.76 | -630.90*** | -6.98*** | -25.96*** |
| | (1.83) | (0.04) | (0.92) | (199.74) | (1.82) | (6.19) |
| Treated × First generation | -3.89* | -0.07 | -3.82* | -440.81** | -1.04 | -2.09 |
| | (1.90) | (0.14) | (1.87) | (193.45) | (1.95) | (2.95) |
| Observations | 1,353 | 1,343 | 1,353 | 1,353 | 1,353 | 1,353 |

*Notes:* This table replicates Table 4 using the exact-match sample, which restricts to randomization blocks in which treated and control instructors taught sections of the same course. Standard errors clustered at the course-section level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: Treatment Effects by Discipline (Exact-Match Sample)

| | (1) Final grades | (2) Course GPA | (3) HW score | (4) Page views | (5) Active days | (6) Total participation |
|---|---|---|---|---|---|---|
| *Panel A. STEM* | | | | | | |
| Treated | -5.77 | -0.11** | -1.25 | -1095.54* | -4.46*** | -28.03** |
| | (3.65) | (0.04) | (1.15) | (409.07) | (0.69) | (9.23) |
| Control SD | 13.73 | 1.08 | 13.77 | 2648.56 | 14.07 | 34.53 |
| Observations | 610 | 603 | 610 | 610 | 610 | 610 |
| *Panel B. Humanities* | | | | | | |
| Treated | -2.81** | -0.12** | -0.72 | -149.44 | -2.55 | -26.86* |
| | (0.96) | (0.05) | (1.14) | (98.11) | (1.44) | (13.04) |
| Control SD | 5.35 | 0.44 | 7.06 | 933.69 | 11.20 | 28.75 |
| Observations | 401 | 399 | 401 | 401 | 401 | 401 |
| *Panel C. Social Sciences/Education/Policy* | | | | | | |
| Treated | -0.59 | 0.03 | -0.92 | -625.18** | -17.55*** | -22.33* |
| | (1.99) | (0.05) | (2.14) | (206.33) | (3.56) | (9.41) |
| Control SD | 8.00 | 0.57 | 8.90 | 1050.19 | 13.85 | 18.77 |
| Observations | 342 | 341 | 342 | 342 | 342 | 342 |

*Notes:* This table replicates Table 5 using the exact-match sample, which restricts to randomization blocks in which treated and control instructors taught sections of the same course. Standard errors clustered at the course-section level are in parentheses. $^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$.

Table A6: Instrumental-Variable Estimates Using Assignment to VSA Access

| | (1) Final grades | (2) Course GPA | (3) HW score | (4) Page views | (5) Active days | (6) Total participation |
|---|---|---|---|---|---|---|
| *Panel A. Full sample* | | | | | | |
| *First-stage: 0.15, F-stat: 50.5* | | | | | | |
| Used VSA | -20.08* | 0.16 | -6.41 | -4136.51*** | -64.64*** | -157.00*** |
| | (12.06) | (0.43) | (5.43) | (1353.33) | (16.18) | (40.36) |
| Observations | 2,379 | 2,362 | 2,379 | 2,379 | 2,379 | 2,379 |
| *Panel B. Exact-match sample* | | | | | | |
| *First-stage: 0.13, F-stat: 16.5* | | | | | | |
| Used VSA | -31.36* | -0.82*** | -11.70 | -5489.58** | -54.76*** | -200.91*** |
| | (18.41) | (0.28) | (7.14) | (2528.69) | (21.12) | (77.29) |
| Observations | 1,353 | 1,343 | 1,353 | 1,353 | 1,353 | 1,353 |

*Notes:* This table reports two-stage least squares estimates in which an indicator for using the VSA at least once during the semester is instrumented with random assignment to VSA access. Each cell is from a separate regression. All specifications include randomization-block fixed effects and student covariates. Standard errors clustered at the course level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A7: Performance of AI Coding on VSA Chats

| Category / summary | Human $n$ | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|---|
| Direct answer | 131 | 91.04 | 93.13 | 92.08 | – |
| Practice/test preparation | 31 | 85.19 | 74.19 | 79.31 | – |
| Navigation/logistics | 24 | 90.00 | 75.00 | 81.82 | – |
| Feedback on an attempt | 8 | 66.67 | 50.00 | 57.14 | – |
| Other | 6 | 38.46 | 83.33 | 52.63 | – |
| | | | | | |
| Macro average | 200 | 74.27 | 75.13 | 72.60 | – |
| Weighted average | 200 | 87.46 | 86.00 | 86.29 | – |
| Micro average / overall | 200 | 86.00 | 86.00 | 86.00 | 86.00 |

*Notes:* This table reports the performance ofAI coding against human coding from 200 examples. All performance values are percentages.