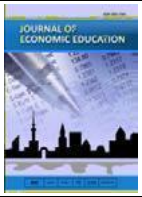
Journal of Economic Education Vol.14 (2) (2025)

Journal of Economic Education

http://journal.unnes.ac.id/sju/index.php/jeec



ALLOWING GENERATIVE AI IN CLASS: EVIDENCE FROM A SEMESTER-LONG CONTROLLED TEACHING STUDY

Christian Rojas,[1]✉ Rong [1], Luke Bloomfield [1]

DOI: https://doi.org/10.15294/jeec.v14i2.34252

[1] Department of Resource Economics, University of Massachusetts Amherst, Massachusetts, United States of America

History Article

Received:
September 24, 2025
Accepted:
October 28, 2025
Published:
December 15, 2025

Keywords:
AI, Teaching, Study

Abstract

We report a controlled, semester-long teaching experiment in an upper-division an- titrust economics course. Two back-to-back sections were held constant in content, assessment, and grading; they differed only in policy and guidance on generative AI: one section was permitted to use AI with disclosure and structured training, the other was prohibited from using AI and received parallel non-AI study guidance (n = 29 vs. n = 28). We find no detectable effect of AI permission on proctored exam scores or final grades. By contrast, AI access is associated with higher engagement on in-class activities, longer and more concentrated AI sessions in other courses (15–30 minutes), greater metacognitive behaviors (preferring one's own answers, catching errors, modi- fying outputs), more positive perceptions—especially regarding efficiency, confidence, and engagement—and stronger intentions to continue using and studying AI, as well as choosing AI-intensive careers. Standardized course evaluations are also consistently higher in the AI section. Taken together, structured AI access with guardrails appears, in our setting, to reshape how students learn and feel about learning, without raising exam scores.

✉ Correspondence address:
181 Presidents Drive, Amherst, MA 01003-9313
E-mail: rojas@umass.edu

p-ISSN 2301-7341
e-ISSN 2502-4485

INTRODUCTION

Generative artificial intelligence (AI) tools, such as ChatGPT, are rapidly diffusing through higher education; yet, credible evidence on how structured access to these tools affects learn- ing, engagement, and the student experience remains limited. Advocates emphasize potential gains in productivity, feedback, and access to examples; critics, however, worry about over- reliance, degraded writing and reasoning, and lack of equity in access (Bond et al., 2024; Vieriu et al., 2025). This paper reports evidence from a semester-long, controlled teaching experiment that isolates the effect of permitting and scaffolding AI use within an economics course where writing and reasoning are crucial components of the curriculum.

The two back-to-back sections of the same upper-division course in Competition (An- titrust) Economics were delivered by the same instructor, on the same days, with identical syllabi, slides, assignments, midterms, grading rubrics, and pacing. The only intended dif- ference was policy and guidance regarding generative AI usage. In the *AI-permitted* section ($n = 29$), students were encouraged to use AI and received structured guidance on effective practices (e.g., brainstorming, revision, summarization) and disclosure norms. In the *non-AI* section ($n = 28$), students were prohibited from using AI and received parallel guidance on traditional study strategies (i.e., note taking, study group, etc.). To make the test conserva- tive, the AI policy was assigned to the afternoon section that historically performs slightly worse. We collected three types of outcomes: (i) **student grades** (two midterms, course grades, homework assignments, and three daily-participation components: iClicker, workout sheets, and exit tickets), (ii) **two survey waves** (midterm and final) focused on AI usage in *other courses* measuring intensity and type of use, perceived helpfulness/effectiveness, depth of engagement (metacognitive behaviors), and forward-looking intentions to engage with AI, and (iii) **standardized university course evaluations** at term end.

Our data analysis suggests three main findings. First, we find no performance pre- mium on exams. Across a suite of incremental regressions, we find no systematic differences between sections on Midterm 1, Midterm 2, or overall course grades; cumulative GPA is the dominant predictor of exam performance. Homework scores are uniformly high in both sec- tions, consistent with ample opportunities to use assistance out of class. On participation components, however, the AI section exhibits substantially higher iClicker performance (sta- tistically significant across specifications), with no detectable differences on workout sheets or exit tickets. We also document two stable patterns independent of AI policy: female students outperform on non-exam components, and first-generation students show stronger participation on several measures.

Second, **AI access changes *how* students use AI outside this course.** Midterm–final surveys (referencing "other courses" to preserve comparability) show that by semester's end the *frequency* distributions of AI use are similar across sections, but the AI section con- centrates use into *longer sessions* (15–30 minutes;

$p < 0.05$ at final). For the AI section, the types of use at the final survey period have a higher tendency to shift toward *edi- torial/grammar* support, while reliance on AI for complete answers remains comparable. Metacognitive behaviors are higher in the AI section: students more often prefer their own answers, catch errors, and modify outputs. Perceived effects move in the same direction and, by the final, there are statistically significantly higher ratings in the AI section for *time efficiency* (homework $p < 0.05$, exam prep $p < 0.01$), *homework grades* ($p < 0.05$), and two skills/dispositions items—*concepts* ($p < 0.05$) and *confidence* ($p < 0.10$).

Third, intentions of future use and course evaluations tilt toward AI. The AI section reports higher forward-looking intentions to use and learn AI across all items, with the *largest, statistically significant* differences on "AI-intensive career" (midterm $p < 0.05$, final $p < 0.01$). Standardized course evaluations favor the AI section on nearly all items, with significant advantages in instructor preparation and use of class time. Students' self-reported expected grades, as recorded in the course evaluations, are similarly distributed across the two sections, alleviating concerns that higher ratings simply reflect grade optimism. By contrast, students in the AI section report lower effort and fewer weekly hours outside of class (consistent with an efficiency channel), alongside higher attendance (consistent with our evidence of greater iClicker participation).

Taken together, the results suggest that structured AI access and guidance do not raise exam scores in this setting but are associated with (i) higher in-class engagement on real-time activities, (ii) deeper, more metacognitive use of AI, (iii) more positive per- ceptions—especially for efficiency and engagement/confidence—and (iv) stronger intentions to continue using and studying AI, including career orientation. In short, the primary ef- fects appear on how students learn and feel about learning, rather than on proctored test performance.

Two caveats are important. First, many outcomes are self-reported; responses may be sensitive to context and demand characteristics ("experimenter effects"): students in the AI-permitted section may answer AI-related items more favorably, while those in the non- AI section may emphasize concerns. We use symmetric wording focused on other courses to mitigate this but cannot eliminate it. Second, the assignment of policy to sections was deliberate (not randomized), and the sample is modest. Therefore, we interpret the estimates cautiously and emphasize patterns that are consistent across multiple measures.

This paper contributes along three margins. First, we provide semester–long, controlled evidence that explicitly varies students' permission to use AI and the structure surrounding that use (explicit guidance and disclosure). The causal effect in our setting, therefore, can be compared with the effect found in a one-shot short-term granting of AI permission. Second, rather than focusing only on test scores, we measure intermediate outcomes that theory and practice identify as plausible AI channels—time/effort, types of use, metacognitive behaviors, perceptions, and forward–looking intentions—thus connecting educational outcomes to the efficiency

and process mechanisms documented in workplace AI experiments (Noy and Zhang, 2023; Brynjolfsson, Li and Raymond, 2023; Peng et al., 2023). Third, our approach to designing AI treatment with guidelines and supporting exercises differs from the earlier literature, where AI use is permitted but no guidelines were provided. Hence, our design helps speak to the effectiveness of pedagogy surrounding AI: whether allowing AI with guardrails correlates with more deliberate, self–regulated use and a more positive classroom climate, consistent with guidance in Kasneci et al. (2023).

Section 2 reviews the literature. Section 3 details the experimental setup. Section 4 reports results on observed performance (exams, grades, participation), midterm/final surveys (intensity, types, metacognition, perceived effects, and outlook), and standardized course evaluations. Section 5 concludes. The Appendix provides the full survey instrument, hand- outs for both AI and non-AI sections, and supplemental figures (including within-AI-section comparisons of "this class" vs. "other courses").

LITERATURE REVIEW

The recent emergence and rapid adoption of generative AI tools, such as ChatGPT, have sparked extensive discussion about their potential impact on educational practices and out- comes. Existing literature explores various dimensions of AI integration in higher education, including academic performance, student engagement, ethical considerations, and pedagog- ical effectiveness.

Research examining AI-enhanced learning outcomes has generally demonstrated positive effects, but the findings remain mixed. For instance, Bommasani et al. (2022) highlights how generative AI tools can significantly augment student productivity and facilitate higher-order thinking by reducing the cognitive load associated with routine tasks. Similarly, Mollick and Mollick (2023) provide empirical evidence indicating improved performance among students permitted to use generative AI tools in various academic contexts, suggesting these technologies can bridge gaps in student understanding when effectively integrated.

However, other studies raise concerns about potential negative consequences of AI tool usage in education, particularly regarding academic integrity and critical thinking skills. Eaton, Crossman and Edino (2023) note an increase in academic misconduct incidents linked to AI-generated submissions, emphasizing the need for clear guidelines and effective pedagog- ical interventions. Similarly, Cotton, Cotton and Shipway (2023) cautions that reliance on generative AI could potentially weaken students' abilities to independently perform essential analytical tasks, highlighting the necessity of balanced instructional design.

Regarding student engagement and study habits, generative AI tools have been shown to alter how students approach learning. Duin and Pedersen (2023) observes that students using generative AI reported higher initial motivation and lower anxiety toward

complex as- signments. Conversely, other scholars have found that students may become passive learners when overly dependent on AI-generated content, potentially undermining intrinsic motiva- tion and engagement (Selwyn, Pangrazio and Cumbo, 2023).

Pedagogical literature underscores the importance of structured guidance in leveraging generative AI tools effectively. According to Zhai, Lin and Chen (2023), explicitly teaching students how to critically engage with AI-generated outputs enhances their ability to dis- cern quality and accuracy, thereby optimizing learning outcomes. Similar findings by Nguyen and McDaniel (2023) stress that merely allowing AI usage without clear instructional frame- works results in inconsistent student performance and engagement, suggesting the need for deliberate integration into curricula.

A complementary line of work highlights a tension between generative AI use and critical thinking. Lee et al. (2025) study knowledge workers and find that task-specific self-confidence correlates with lower confidence in AI outputs and more engaged critical thinking, whereas high confidence in AI outputs is associated with reduced scrutiny. They argue that gen- erative AI can reshape cognitive work by shifting effort from information gathering and problem solving toward verification, output integration, and isolated task completion. This perspective motivates our focus on metacognitive outcomes and the scaffolding we provide (verification, modification, attribution), and helps interpret our findings that AI-permitted students more often prefer their own answers, catch errors, and modify outputs.

Despite extensive theoretical and anecdotal discourse, robust empirical evidence from controlled experiments comparing AI-integrated and traditional instructional methodologies remains limited, and the few existing studies primarily utilize short-term experimental de- signs or observational approaches. Most experiments in the existing literature typically span periods of a few weeks at most, making it difficult to capture longer-term shifts in student behavior and learning outcomes (Mollick and Mollick, 2023). To address this critical gap, the present study makes a unique contribution by providing rigorous empirical insights through a semester-long controlled experiment. This extended timeframe enables deeper and more reliable insights into how sustained AI usage impacts student performance, study habits, and engagement over a substantial academic period.

In summary, while the literature on generative AI in education broadly acknowledges its potential benefits and pitfalls, there is a notable shortage of long-term, experimentally rigorous studies. This paper directly addresses this shortcoming by systematically assessing the impacts of structured generative AI integration over an entire semester, thereby providing valuable empirical evidence for educators and policymakers navigating this rapidly evolving educational landscape.

METHODS

The experiment was implemented in a semester-long, upper-level undergraduate course in Competition (Antitrust) Economics. Two sections of the course were offered on the same days (both Tuesday and Thursday), in close succession: one meeting from 10:00–11:15am and the other from 1:00–2:15pm. The design intentionally assigned the *AI-permitted* policy to the afternoon (1:00pm) section and the *non-AI* policy to the morning (10:00am) section. This assignment was based on the instructor's prior teaching experience: historically, students in the 10:00am section tended to perform slightly better on average. By assigning the AI policy to the section that typically underperforms, the design stacks the odds against detecting spurious AI effects, providing a more conservative test of the impact of AI integration.

We worked with the scheduling office to ensure a similar number of students enrolled in each section; the AI class had 29 students, while the non-AI class had 28 students. Both sections were identical in every respect, except for the AI policy and related guidance. The following elements were held constant across the two courses:

• Course structure: Both sections followed the same syllabus, lecture slides, daily activities (in class and online), and homework assignments (five total).4 Two midterm examinations were administered in a traditional paper-and-pencil format, with no notes, cheatsheets, technology, or AI permitted. Both exams used identical questions and grading rubrics across sections. Course grades were determined by exams (60%),5 homework (20%), and in-class activities (20%).
• Surveys: Both sections completed two comprehensive surveys during the semester ("survey" henceforth); the full instrument is included in Appendix Table A1. The survey elicited (i) frequency and duration of AI use, (ii) types of use, (iii) depth of engagement/metacognitive behaviors, (iv) perceived effects (helpfulness/harmfulness) on skills, time, and performance, and (v) forward-looking intentions (continued use, learning, and career orientation). To enhance comparability and limit demand effects, most items asked about behavior in other college courses (i.e., excluding this class). In addition, the AI section answered a short module about AI use in this class; the non-AI section received a parallel module that replaced references to AI with "class notes/materials."

• Course evaluations: At the end of the semester, both sections completed the standardized university course evaluation ("evaluation" henceforth), which provides an independent measure of teaching quality and student experience.

The only intentional differences between the sections were as follows:

(a) AI policy: In the non-AI section, the syllabus specified: “Using generative AI tools as a substitute for your authentic intellectual and creative work is considered cheating. In other words, the work you submit must be your own.” In the AI section, the syllabus instead encouraged students to use generative AI to support their learning and course-work efficiency, while also requiring disclosure of AI use in assignments (e.g., attribution of generated text, use of colored fonts, quotations, or parenthetical citation): “You can choose to use generative AI tools to support your learning and to be more efficient in your assignments. For example, you can use AI to help brainstorm assignments or projects, to revise existing work you have written, to summarize/organize information in class notes or other documents. I will provide some tools that can help you make an efficient use of AI to support your learning. When you submit your work, I expect you to clearly attribute what text was generated by the AI tool (e.g., AI-generated text appears in a different colored font, quoted directly in the text, in-text parenthetical cita- tion, or a disclosure that provides details of how you used generative AI). I will provide details of what is required in terms of AI use disclosure for each assignment.”
(b) Guidance lecture:In the AI section, half of the first lecture was devoted to struc- tured guidance on effective AI use for brainstorming, revision, summarization, and information organization, accompanied by a handout (see Appendix C). In the non-AI section, the same amount of time was dedicated to strategies for studying and com- pleting assignments without AI, with a parallel handout provided. In both sections, students were reminded throughout the semester of the course policy and corresponding guidance (AI use in the AI section, and alternative learning resources in the non-AI section).
(c) Survey adaptation: As noted above, survey questions directly referencing AI in “this class” were reworded for the non-AI section (Appendix Table A1).

A summary of similarity and differences between the two classes is shown in Table 1. To assess comparability, we examined the baseline characteristics of the two sections (see Table 2). Overall, the classes were fairly well balanced in terms of enrollment, demographics, and prior academic performance, though minor differences remain, as is typical in field settings.

Table 1. Comparison of Course Setup Across AI and Non-AI Sections

Feature	AI Section (1:00–2:15pm)	Non-AI Section (10:00–11:15am)
Common Elements		
Course content	Identical syllabus, lecture slides, daily activities, homework (5), and midterms (2)	
Grading scheme	Exams 60%[a], Homework 20%, In-class activities 20%	
Surveys	Two extensive surveys during the semester (Appendix Table A1)	
Course evaluation	Standard university evaluation at semester end	
Differences		
Policy on AI	Encouraged AI use with disclosure requirements	AI use prohibited (considered cheating)
Guidance lecture	Half lecture devoted to strategies for effective AI use; AI handout (Appendix C)	Half lecture on non-AI study/test-prep strategies; non-AI handout (Appendix C)
Survey wording	Questions referred directly to AI use in "this class"	Questions reworded to remove AI references in "this class"
Assignment disclosure	Attribution of AI use required (colored font, quotation, citation, or disclosure)	Not applicable (AI prohibited)

[a] Students could reweight midterms (60/40 vs. 50/50) in a way that only improved their grade.

Our analysis therefore focuses on whether allowing AI access causally affected three dimensions of student outcomes: (i) observed performance (midterm scores and overall course grades), (ii) survey responses on AI perceptions, usage, and effects, and (iii) standardized course evaluations.

Table 2. Balance Table: AI vs. Non-AI Class

Variable	AI Class (% or Mean)	Non-AI Class (% or Mean)	p-value
Cum GPA	3.25	3.11	0.272
Sex: Male	62.1%	60.7%	1.000
Sex: Female	37.9%	39.3%	
First Gen: No	72.4%	67.9%	0.777
First Gen: Yes	27.6%	32.1%	
Ethnicity: White	24.1%	32.1%	0.903
Ethnicity: Asian	51.7%	39.3%	
Ethnicity: Black	6.9%	7.1%	
Ethnicity: Hispanic	6.9%	10.7%	

Ethnicity: Multiple	3.4%	3.6%	
Entrance: Freshman	75.9%	50.0%	0.057
Entrance: Transfer	24.1%	50.0%	
Citizenship: International	31.0%	32.1%	0.913
Citizenship: Permanent Res.	10.3%	7.1%	
Citizenship: US Citizen	58.6%	60.7%	
Residency: In-State	48.3%	53.6%	0.903
Residency: Int'l	31.0%	28.6%	
Residency: Out-of-State	20.7%	17.9%	

Notes: Table compares baseline characteristics of students in the AI-allowed section ($n = 29$) and the non-AI section ($n = 28$). Cum GPA is continuous and tested using Welch's *t*-test. Categorical variables are tested using Fisher's exact test (for binary variables) or chi-squared test of independence (for variables with more than two categories). For brevity, p-values are reported only once per categorical variable group

RESULT AND DISCUSSION

Observed Performance

Tables 3 and 4 present a series of incremental regressions, where additional covariates are sequentially introduced from left to right. These models conduct OLS regressions to test whether students in the AI section exhibited differential average performance on the midterms. For Midterm 1, Model 1 suggests a positive and somewhat sizable effect of being in the AI section (about five points), although the coefficient declines in magnitude and loses statistical significance once controls are added. It is worth noting, however, that the absence of statistical significance may reflect limited statistical power rather than the absence of a true effect, given the relatively small sample size. By contrast, results for Midterm 2 show no evidence of an AI effect across specifications. Overall, the most robust and consistent predictor of exam performance is students' prior academic achievement, as measured by cumulative GPA.

Table 3. Incremental Regression Results for Midterm 1 Grade

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
AI	5.32*	3.66	3.66	3.54	2.79	1.61
	(2.89)	(2.52)	(2.55)	(2.58)	(2.67)	(2.72)
Cum GPA		12.65***	12.65***	12.08***	11.51***	11.94***
		(2.89)	(2.91)	(3.36)	(3.69)	(3.60)
Female			0.02	-0.85	-0.51	-0.80
			(2.62)	(2.75)	(3.10)	(3.03)
Int'l				2.37	2.17	1.87
				(3.31)	(3.97)	(4.10)
Out of State				3.90	3.03	4.97
				(3.70)	(4.11)	(4.12)
Black					-1.65	-4.94

					(5.44)	(5.54)
Hispanic					5.16	2.62
					(4.99)	(5.02)
Transfer						-5.58
						(3.33)
First Gen.						3.37
						(3.00)
Intercept	78.56***	39.15***	39.14***	39.98***	39.62***	41.54***
	(2.06)	(9.16)	(9.30)	(10.36)	(11.58)	(11.53)
R^2	0.06	0.31	0.31	0.33	0.37	0.43
N (AI / non-AI)	28 / 27	28 / 27	28 / 27	28 / 27	27 / 25	27 / 25

Notes: OLS regressions, Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

No differential effect is observed in the OLS regression analysis when examining overall course grades (out of 100 points), which incorporate both exam and non-exam components (Table 5). A noteworthy feature of these results, however, is that female students outperform their male counterparts by a statistically significant margin (up to seven points). Impor- tantly, this advantage appears to stem from non-exam components, as no comparable gender difference was detected in the midterm regressions.

Table 4. Incremental Regression Results for Midterm 2 Grade

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
AI	1.36	-0.20	-0.20	0.04	-0.11	-1.98
	(2.90)	(2.41)	(2.44)	(2.42)	(2.55)	(2.46)
Cum GPA		13.78***	13.78***	11.87***	12.21***	12.33***
		(2.72)	(2.75)	(3.10)	(3.42)	(3.19)
Female			-0.20	-1.86	-0.85	-0.86
			(2.51)	(2.62)	(3.02)	(2.81)
Int'l				5.14	6.09	2.95
				(3.12)	(3.65)	(3.55)
Out of State				4.96	3.64	4.82
				(3.49)	(3.88)	(3.67)
Black					-1.80	-3.87
					(5.67)	(5.37)
Hispanic					4.95	2.58
					(4.38)	(4.22)
Transfer						-7.81**
						(3.03)
First Gen.						-3.82
						(2.61)
Intercept	85.85***	43.01***	43.09***	47.10***	44.43***	50.86***
	(2.11)	(8.63)	(8.77)	(9.50)	(10.68)	(10.16)

R^2	0.00	0.33	0.33	0.38	0.43	0.53
N (AI / non-AI)	29 / 26	29 / 26	29 / 26	29 / 26	28 / 24	28 / 24

Notes: OLS regressions, Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

We next examined differences in the non-exam components of the course in greater detail. Homework performance was uniformly high across both sections. This outcome is unsurpris- ing given the essay-based nature of the assignments and the fact that, in practice, students in both sections had access to AI tools outside of class (making AI use likely regardless of the official policy). Consequently, no meaningful performance gap emerged between the AI and non-AI groups for homework.

Table 5. Incremental Regression Results for Final Grade

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
AI	1.29	-0.74	-0.42	-0.51	-0.56	-1.30
	(3.49)	(2.82)	(2.72)	(2.73)	(2.56)	(2.44)
Cum GPA		18.09***	17.97***	17.23***	16.92***	17.79***
		(3.23)	(3.11)	(3.54) 5.31*	(3.52)	(3.21)
Female			6.48**	(2.94)	7.21**	6.66**
			(2.79)		(2.99)	(2.73)
Int'l				2.96	0.57	1.20
				(3.51)	(3.80)	(3.63)
Out of State				5.10	1.00	3.58
				(3.98)	(3.97)	(3.72)
Black					-1.92	-6.22
					(5.32)	(5.05)
Hispanic					4.12	1.30
					(4.58)	(4.32)
Transfer First						-5.45*
						(3.03)
Gen.						7.31***
						(2.64)
Intercept	81.26***	24.99**	22.83**	23.90**	25.24**	24.57**
	(2.49)	(10.24)	(9.89)	(10.87)	(11.06)	(10.29)
R^2	0.00	0.37	0.43	0.45	0.60	0.69
N (AI / non-AI)	29 / 28	29 / 28	29 / 28	29 / 28	28 / 26	28 / 26

Notes: OLS regressions, Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Daily activities consisted of three types: two conducted in class and one completed online outside of class. The first in-class activity required students to respond in real time to prompts displayed on the lecture slides using an electronic response system (iClicker). The second involved "workout sheets" in which students solved a problem—individually or in groups—directly related to the concept covered that day. The third, completed

outside of class, required students to answer a short prompt on the university's learning management system based on that day's lecture; this activity was labeled the "exit ticket."

We ran OLS regressions to examine the impact of AI-policy across these three types of daily activities (see Tables 6, 7, and 8). The results indicate a significant effect of AI policy only for the iClicker activities, with students in the AI section showing substantially higher performance in the in-class iClicker sessions. By contrast, no evidence of an AI effect was detected for either the in-class workout sheets or the take-home exit tickets. It is worth noting, however, that the estimated coefficients for AI are consistently positive and nontrivial in magnitude across specifications. The absence of statistical significance here may therefore reflect limited statistical power rather than the absence of a meaningful effect, a caveat already noted in the discussion of exam results.

Two additional patterns are noteworthy. First, female students consistently outperformed their male peers on the workout sheets and exit tickets, but not on the iClicker tasks. Second, first-generation students registered markedly higher scores on two of the three participation measures (workout sheets and exit tickets), suggesting elevated levels of engagement in these components relative to their peers.

Midterm and Final Survei

We report results of the survey we crafted for the two sections, which was distributed (in identical format) at the time of the first midterm (middle of the semester) and at the time of the second midterm (last week of the semester). We refer to these two time periods as "Midterm" and "Final". The analysis below compares the responses of the AI section to those of the non-AI section, separately for the midterm and final periods. The questions ask about AI use and perceptions as they pertain to "other classes"; this ensures we can have comparability across the two sections (since the non-AI class could not, by design, use AI in this class). For the AI section students, however, we had a separate set of questions, where we ask about their AI use and perceptions for this class; Appendix Tables A1, A2, A3 and A4 compare within–class vs. other–class usage for the AI section.

Intensity of Use

We measure students' intensity of generative–AI use outside this course along two margins: (i) *frequency* of use for assignments in other college courses and (ii) *minutes per assignment when AI is used (both items exclude the present class).*

Table 6. iClicker Regression Results by Model Specification

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
AI	22.82***	19.37***	19.74***	18.68**	15.44**	16.82**
	(7.26)	(7.00)	(7.04)	(7.17)	(6.92)	(7.37)
Cum GPA		21.49**	21.74**	25.26**	35.29***	35.82***
		(8.30)	(8.33)	(9.60)	(9.98)	(10.17)
Female			5.93	5.60	7.99	7.46

			(7.05)	(7.57)	(7.82)	(7.79)
Int'l				-4.34	-10.41	-5.75
				(9.04)	(9.88)	(10.45)
Out of State				8.11	4.01	6.61
				(10.53)	(10.79)	(11.19)
Black					-1.39	-2.49
					(13.69)	(13.98)
Hispanic					38.24***	40.80***
					(12.79)	(13.01)
Transfer						4.48
						(8.95)
First Gen.						11.68
						(7.68)
Intercept	59.30***	-7.64	-11.05	-21.45	-58.93*	-67.50**
	(5.23)	(26.33)	(26.72)	(29.70)	(32.10)	(32.39)
R-squared	0.165	0.266	0.276	0.294	0.457	0.491
N (AI / non-AI)	27 / 25	27 / 25	27 / 25	27 / 25	26 / 24	26 / 24

Notes: OLS regressions, Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 7. Work Out Sheet Participation Regression Results by Model Specification

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
AI	7.54	3.70	4.22	4.60	5.59	7.71
	(6.07)	(5.54)	(5.15)	(5.21)	(5.50)	(5.18)
Cum GPA		25.01***	25.37***	21.60***	20.03** (8.16)	21.84***
		(6.68)	(6.20)	(7.14) 14.15**	12.00*	(7.36) 11.75**
Female			16.06***	(5.52)	(6.26)	(5.57)
			(5.23)			
Int'l				8.13	3.97	10.49
				(6.58)	(8.04)	(7.54)
Out of State				3.61	4.41	6.67
				(7.48)	(8.34)	(7.66)
Black					5.27	1.79
					(11.22)	(10.31)
Hispanic					-6.03	-6.43
					(9.85)	(8.93)
Transfer						4.15
						(6.35)
First Gen.						19.71***
						(5.42)
Intercept	68.65***	-9.13	-16.56	-7.22	-1.74	-16.20
	(4.29)	(21.13)	(19.77)	(22.00)	(26.04)	(23.42)
R-squared	0.028	0.231	0.349	0.369	0.425	0.569
N (AI / non-AI)	28 / 28	28 / 28	28 / 28	28 / 28	27 / 26	27 / 26

Notes: OLS regressions, Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 8. Exit Ticket Participation Regression Results by Model Specification

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
AI	6.53	4.25	4.74	4.51	4.00	4.98
	(5.41)	(4.94)	(4.82)	(4.78)	(4.58)	(4.38)
Cum GPA		20.33***	20.14***	18.59***	19.65***	20.78***
		(5.65)	(5.50)	(6.19)	(6.30)	(5.76)
Female			9.87*	7.20	10.19*	9.72*
			(4.95)	(5.14)	(5.35)	(4.89)
Int'l				6.56	2.85	7.18
				(6.14)	(6.79)	(6.51)
Out of State				11.83*	4.70	7.38
				(6.97)	(7.11)	(6.67)
Black					-3.43	-8.01
					(9.51)	(9.05)
Hispanic					1.23	-0.83
					(8.19)	(7.73)
Transfer						-0.56
						(5.44)
First Gen.						15.64***
						(4.74)
Intercept	69.03***	5.81	2.52	4.61	5.58	-1.91
	(3.86)	(17.92)	(17.52)	(19.01)	(19.79)	(18.44)
R-squared	0.026	0.214	0.269	0.313	0.464	0.575
N (AI / non-AI)	29 / 28	29 / 28	29 / 28	29 / 28	28 / 26	28 / 26

Notes: OLS regressions, Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figure 1 shows the distribution of self–reported frequency. At midterm, students in the non–AI section are more likely to report never using AI in their other courses (about one–third vs. about one–tenth in the AI section; Fisher's exact test $p < 0.10$ for the "Never" bin). By the final, the two distributions converge: differences across bins are small, and none of the comparisons are statistically significant at conventional levels.

Figure 2 reports the distribution of time spent with AI per assignment. At midterm, the two sections look similar, with a common mode at 15–30 minutes and no bin–level differences. By the end of the term, however, AI–section students are substantially more likely to report spending 15–30 minutes with AI (roughly one–half, vs. roughly one–quarter in the non–AI section; Fisher's exact test $p < 0.05$ for the 15–30 minute bin). No other bin differs significantly.

These patterns suggest that access to structured AI guidance does not make students use AI more frequently in their other courses by semester's end, but it does shift usage toward longer, more substantive sessions when they do engage. This is consistent with an efficiency story in which students concentrate AI use into focused bouts rather

than frequent, shallow interactions.

Types of Use

We classify reported uses of generative AI in other college courses (excluding this class) into five categories: *editorial* (grammar/editing), *examples* (obtaining additional examples), *clarify* (clarifying essential concepts), *answer* (obtaining a sample answer), and *other*. For each category, we collapse the five-point frequency scale into an indicator for *regular use* (selected "About half the time," "Most of the time," or "Always"). Figure 3 plots, by section, the fraction of students who meet this threshold at midterm and final.

At midterm, the AI section reports higher regular use in all categories, though none of the bin-level gaps are statistically significant at conventional levels. By the end of the semester, a marked difference emerges in editorial use: a substantially larger share of AI section students reports using AI for grammar/editing at least half the time ($p < 0.05$). In contrast, clarify converges across sections, examples remain similar to midterm in relative magnitude, answer shows a persistent but statistically indistinct gap, and other widens modestly in favor of the AI section.

In sum, these results suggest that structured exposure and guidance are associated with a shift toward revision-oriented use (editing/rewriting) rather than increased reliance on AI for full answers, and that concept clarification becomes widespread in both groups by the end of the term.

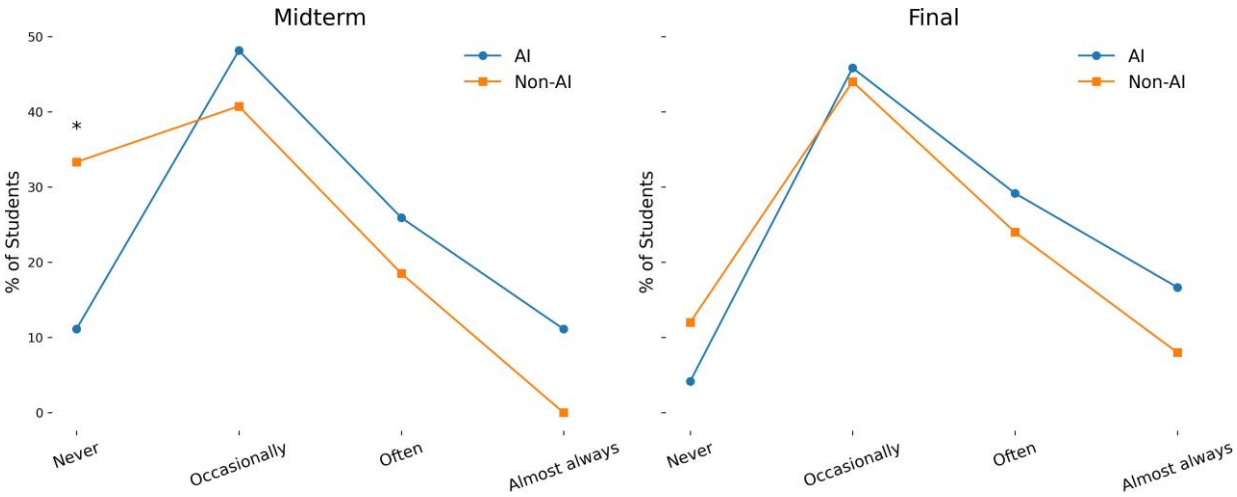


Figure 1. How Often Students Use Generative AI in *Other* Courses (Midterm vs. Final)

Depth of Engagement and Metacognitive Use

We also probe how students interact with generative AI at a deeper, more metacognitive level in their *other* college courses (excluding this class). Specifically, when they use AI for an assignment, how often do they: (i) *prefer their own answer* over the AI's, (ii) *catch a mistake* in the AI output, (iii) *modify* the AI output to fit

their needs, (iv) *learn something new* from the AI output, and (v) find the AI output *unclear/not helpful*. For each item we collapse responses to an indicator for *regular use* (reporting "About half the time," "Most of the time," or "Always").

Figure 4 plots the share of students in each section who meet this threshold at midterm and at the end of the semester. At midterm, the AI section exhibits higher regular engage- ment on four of five behaviors; the largest gap is for *preferring one's own answer* (AI > Non–AI; Fisher's exact test $p < 0.10$). By the final, gaps are broad and consistent—AI- section students more often report *preferring their own answer*, *catching errors*, *modifying* model output, and *learning from AI*—while the incidence of *unclear/not helpful* outputs remain low and comparable across sections. Although end-of-semester binwise differences are not statistically significant at conventional levels (reflecting small samples and multiple bin tests), several effect sizes are large (on the order of 15–20 percentage points).

The observed patterns point to more *metacognitive* AI use among students who received structured guidance: they scrutinize, adapt, and at times override AI outputs. They also report learning something new more often, rather than simply accepting the AI's answers.

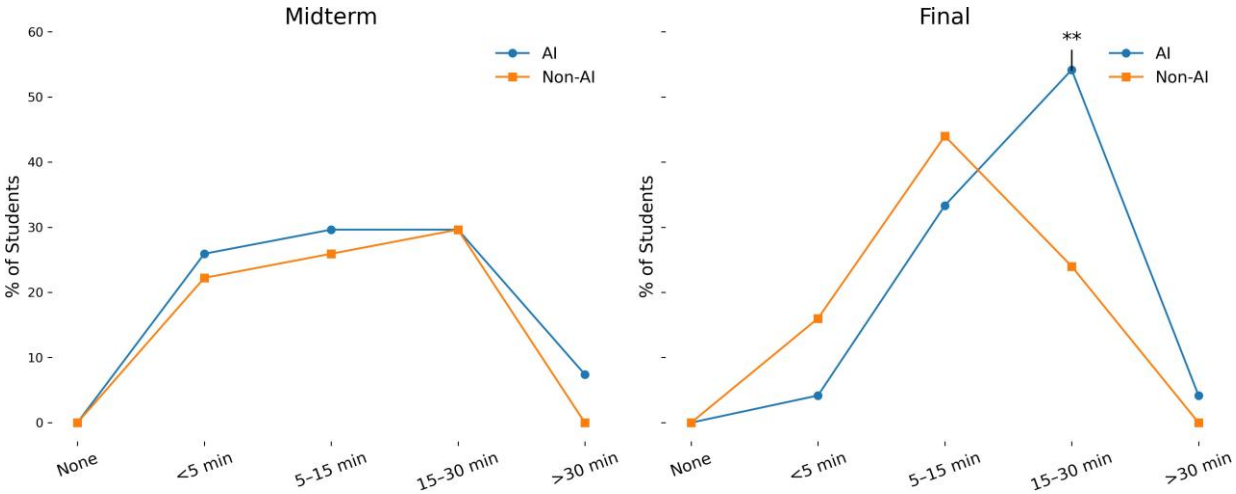


Figure 2. Minutes Spent with Generative AI per Assignment in *Other* Courses (Midterm vs. Final)

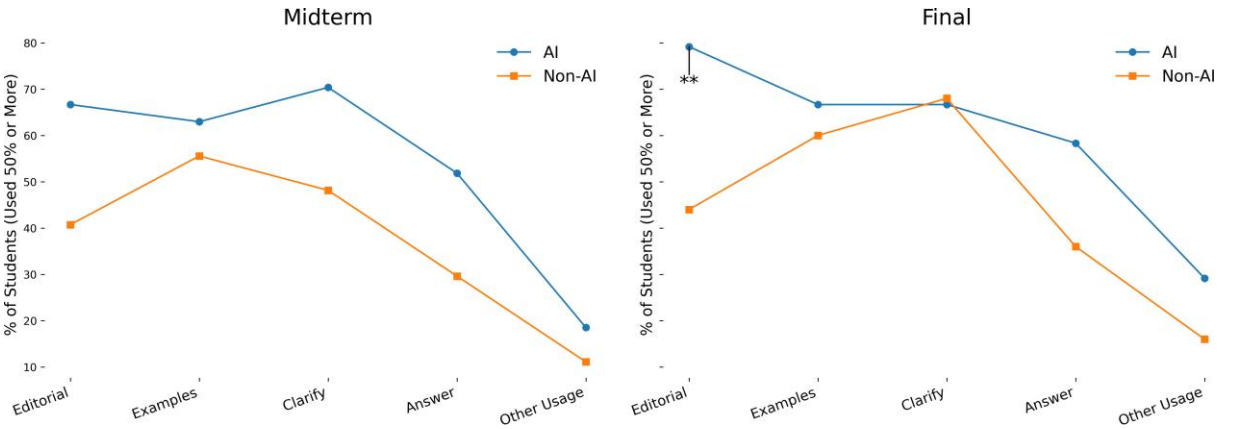


Figure 3. Types of Generative-AI Use in *Other* Courses: Share Using the Mode ≥50% of the Time

Perceived Effects of AI on Skills, Dispositions, Time, and Performance

Students rated how *helpful or harmful* generative AI is across two sets of outcomes using a −10 to 10 scale (−10 = very harmful, 0 = neutral, 10 = very helpful). The first set captures skills/dispositions: understanding key concepts, writing quality, creativity, sense of ownership of one's work, work ethics/morality, confidence in the subject, and engagement with the material. The second set covers time and performance: time to complete homework, time to prepare for exams, time spent on overall coursework, homework grades, overall course grade, and overall learning experience.

Figure 5 plots section means for the first set at midterm and final. Two patterns stand out. First, at both waves the AI section reports higher values on every item. Second, these differences widen by the final: AI students rate concepts, confidence, and engagement roughly 1.5–3 points higher (on the −10 to 10 scale) than non-AI students, and their views on creativity move from near zero at midterm to clearly positive at final. Wilcoxon rank-sum tests (two-sided) indicate that, by the final, the AI section is significantly higher on Concepts ($p < 0.05$) and Confidence ($p < 0.10$); (no midterm differences reach $p < 0.10$). Perceptions on ownership and ethics remain net negative overall but are less negative in the AI section at both waves, though these differences are not statistically significant.

Figure 6 shows the second set. The AI section is consistently more optimistic, and the gap widens by the end of the semester. By the final, Wilcoxon rank–sum tests (two-sided) show significantly greater perceived help for reducing time on homework ($p < 0.05$) and exam prep ($p < 0.01$), and for homework grades ($p < 0.05$); means on these items are roughly 5–6 for the AI section versus about 3 for the non-AI section. A significant midterm advantage on Course Grade ($p < 0.01$) attenuates and is not significant by the final.

Overall, students with structured AI access report increasingly positive views over the term. The largest and most reliable gains concentrate in efficiency (less time for home- work and exam prep) and affective/disposition dimensions—especially confidence (and directionally, engagement and concepts). Perceived performance effects are mixed: homework grades show an advantage by the final, while a midterm bump in course grade dissipates by semester's end; overall learning is higher on average but not significant. Concerns around ownership/ethics persist but are less negative in the AI section. Given modest sample sizes, some nulls may reflect limited power.

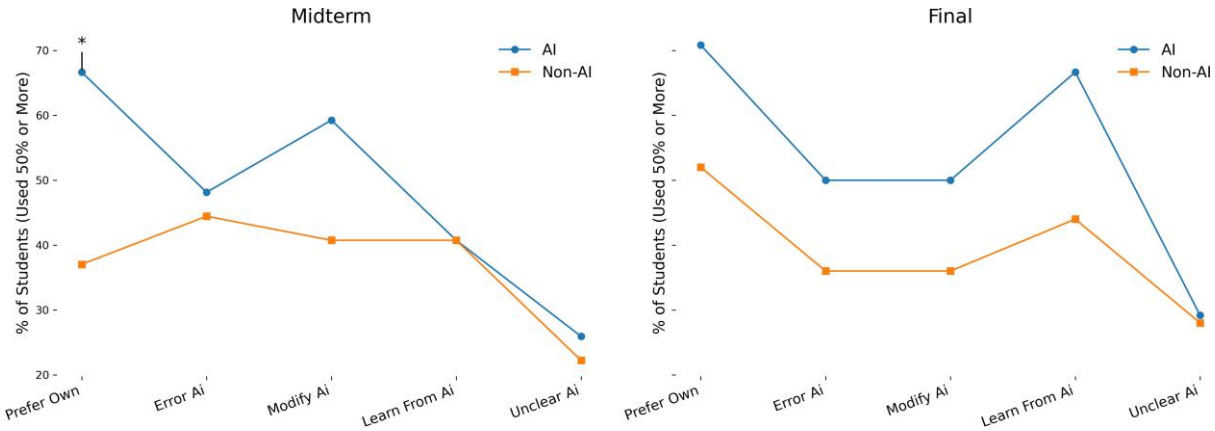


Figure 4. Metacognitive Behaviors When Using Generative AI in *Other* Courses: Share Reporting the Behavior ≥50% of the Time

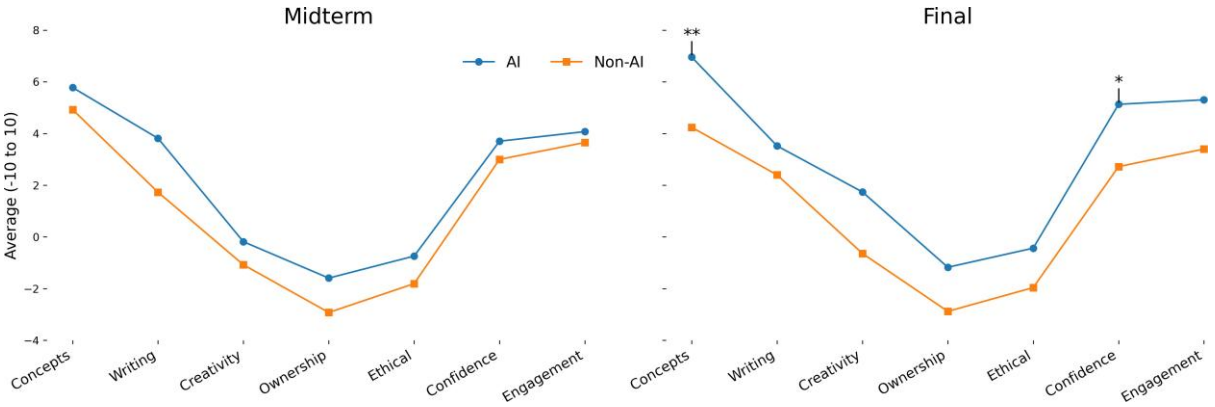


Figure 5. Perceived Effects of Generative AI on Skills and Dispositions (Mean, −10 to 10)

AI Adoption Intentions and Career Orientation

Students rated their forward-looking intentions toward AI on a 0–10 scale (0 = not likely, 10 = extremely likely): (i) *use AI in future classes*, (ii) *recommend AI for coursework to friends,* (iii) *take more courses with AI integration*, (iv) *spend significant time learning how to use AI efficiently*, and (v) *choose a career that requires extensive AI use.*

Figure 7 shows section means at midterm and at the end of the semester. Two patterns are clear. First, at both waves, the AI section reports *higher* intended adoption across all five items. Second, the gap tends to widen by the final—especially for *AI career*, where the difference is statistically significant at midterm ($p < 0.05$) and grows to a larger, highly significant difference at the final ($p < 0.01$). Other items display similar directional gaps (roughly 1–2 points on the 0–10 scale) but are not statistically significant given the sample size.

Within-section changes over time are also informative: mean intentions rise for both sections, but the increase is larger in the AI section, notably for *taking more AI courses* and *future AI use*. We interpret this as consistent with structured exposure

increasing students' interest, confidence, and motivation to deepen their AI skills and, for some, to see AI as part of their career trajectory.

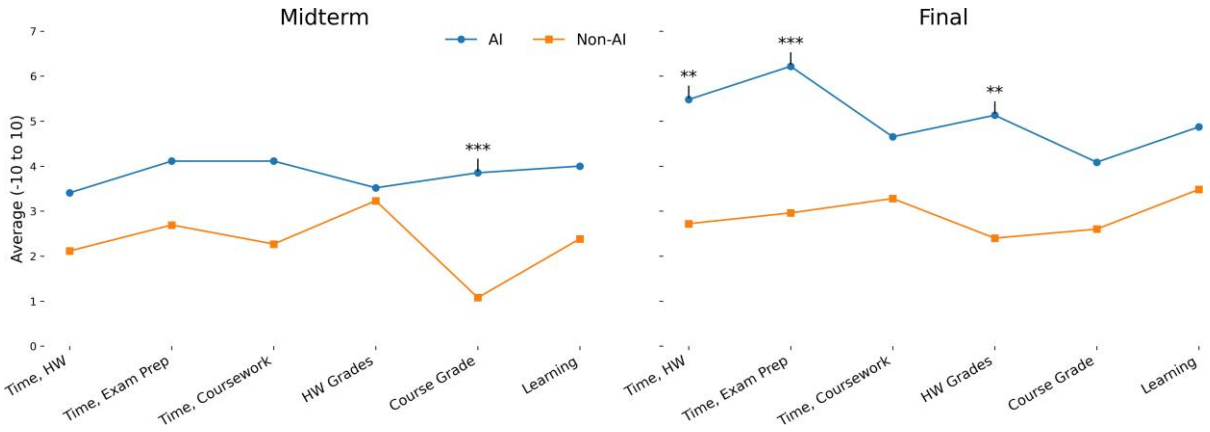


Figure 6. Perceived Effects of Generative AI on Time and Performance (Mean, −10 to 10)

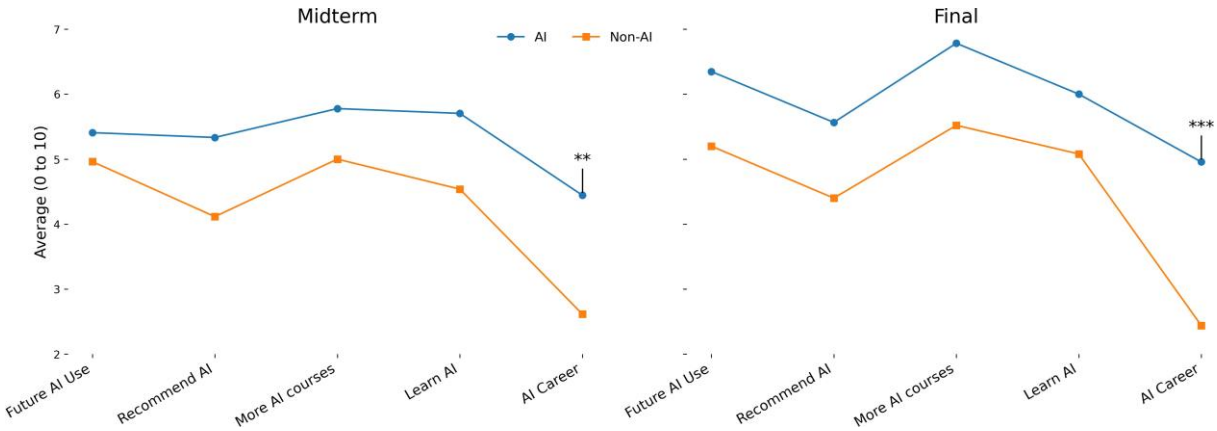


Figure 7. Intentions to Use and Engage with AI (Mean Likelihood, 0–10)

Survey takeaways

Across the midterm and final surveys, a coherent pattern emerges. By semester's end, the two sections report similar *frequency* of AI use in their other courses, but students in the AI section concentrate that use into longer, more substantive sessions (15–30 minutes). They also display greater *metacognitive* engagement—preferring their own answers, detecting errors, and modifying outputs—with gaps that widen by the final. Perceived effects of AI are consistently more favorable in the AI section in both survey waves and become more positive over time, particularly for *efficiency* (less time on homework and exam preparation) and for *engagement/confidence*, while concerns about ownership/ethics are less negative. Looking ahead, AI-section students express stronger adoption intentions across all

items, with the largest and statistically significant differences for pursuing an AI-intensive career (midterm $p < 0.05$; final $p < 0.01$). In short, structured guidance appears to shift students toward deeper, more deliberate AI use, more positive attitudes, and stronger intentions to continue using and studying AI, without increasing raw frequency of use.

We conclude this summary, with an important caveat. These findings are based on self- reported survey measures and may be sensitive to "experimenter effects". For example, students may answer AI-related items more favorably when their instructor has permitted and scaffolded AI use, whereas students in the non-AI section may understate benefits or emphasize concerns; interpretations should account for this potential response bias.

Standardized Course Evaluation

This section reports results from the university's standardized end-of-semester course eval- uation survey. The instrument consists of twelve questions designed to assess students' perceptions of both instructional effectiveness and the overall value of the course. Responses are recorded on a five-point Likert scale, where 1 denotes "one of the worst" and 5 denotes "one of the best." Figure 8 plots the mean response for each item, separately for the AI and non-AI sections.

The most consistent pattern is that the AI section reports higher average ratings across all twelve items (with the exception of Q12, where the two sections are tied). Although differences are generally modest in magnitude, two questions—Q1 (instructor preparation) and Q4 (instructor's use of class time)—show statistically significant differences at a 10% level. Overall, the evidence indicates that students in the AI section evaluated both the instructor and the course more positively than their peers in the non-AI section.

The survey also elicited information on students' expected course grade (Figure 9), self- reported effort (Figure 10), time spent outside the classroom (Figure 11), and attendance (Figure 12). Notably, the more favorable evaluations of the AI section (noted in Figure 8) cannot be attributed to inflated grade expectations. As shown in Figure 9, the distribution of anticipated course grades is highly similar across the two sections, consistent with the performance patterns observed in the midterm and final grade analyses presented earlier. Both sections report comparable shares of students expecting an A or A–, but the AI section has a noticeably larger share in the A– category and a slightly smaller share expecting a straight A. This pattern suggests, if anything, that students in the AI section anticipated slightly lower outcomes, making it unlikely that the higher evaluation ratings reflect grade optimism or grade inflation.

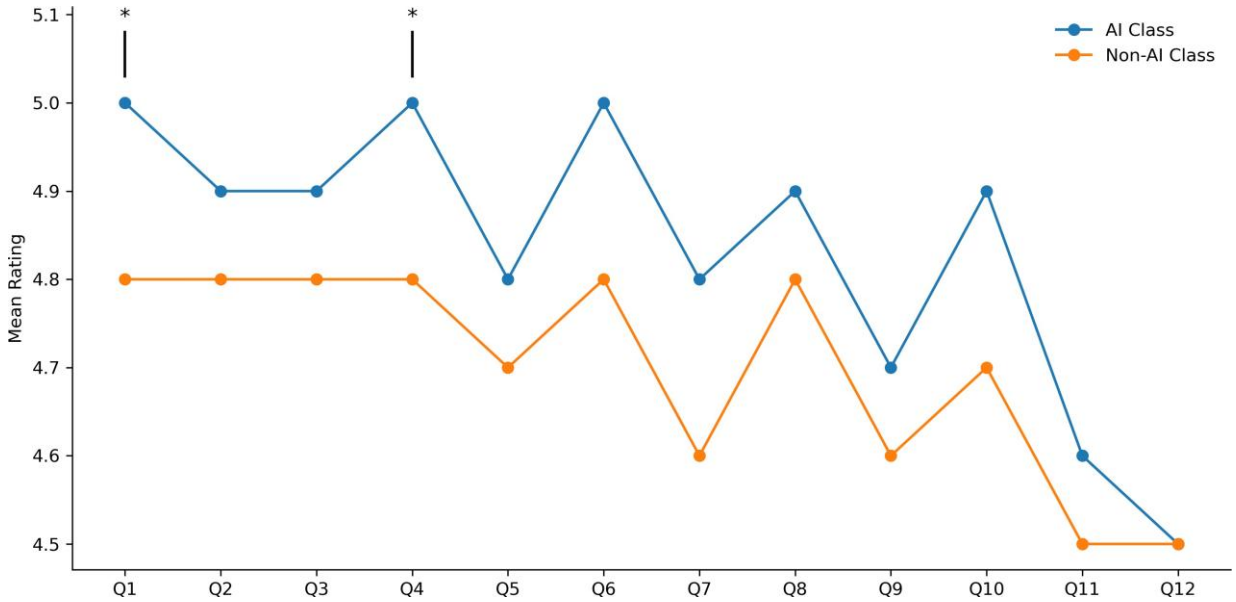


Figure 8. Comparison of Mean Ratings Across Q1–Q12 Between AI and Non-AI Classes

The results in Figures 10 and 11 suggest that access to AI tools primarily operated through an efficiency channel: students in the AI section reported both lower overall effort levels and fewer hours spent working outside of class.

Regarding effort (Figure 10), the two largest categories— "medium" and "high"—account for roughly 90% of students in both sections. In the non-AI section, students were evenly split between these two categories (45% each). By contrast, in the AI section a substantially larger proportion of students reported "medium" effort (52%) relative to "high" effort (36%), with a small share (8%) indicating "very high" effort.

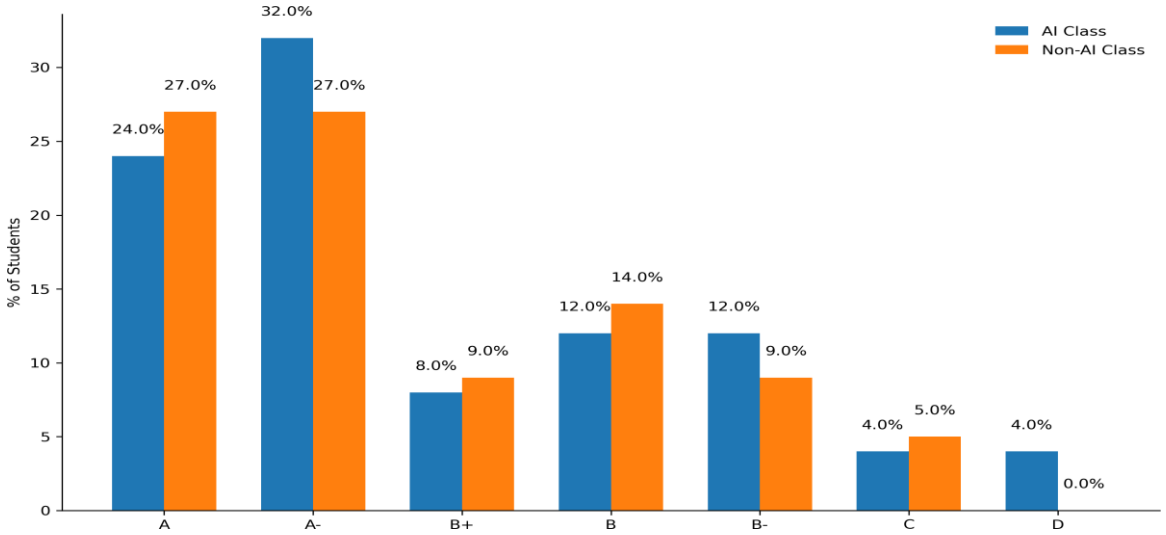


Figure 9. Distribution of Expected Course Grades in AI and Non-AI Sections

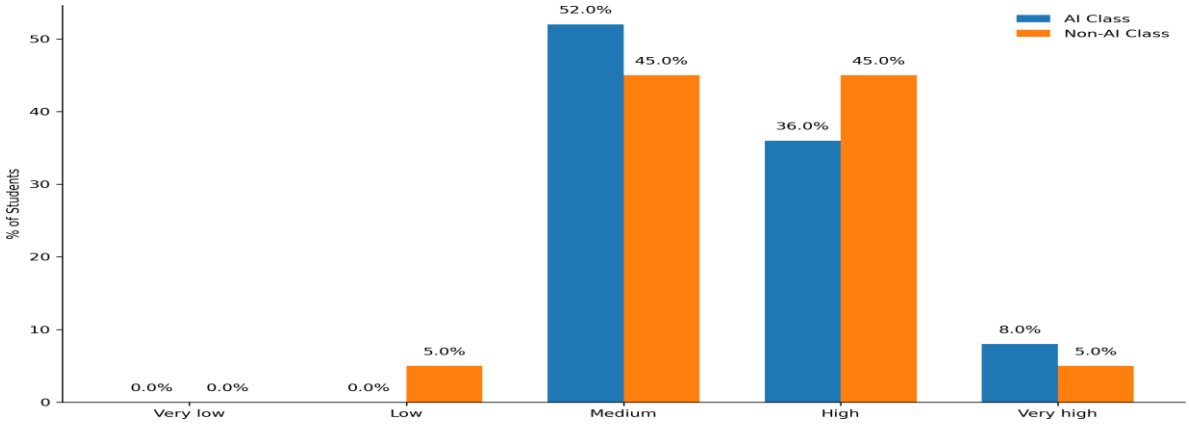


Figure 10. Student-Reported Effort Levels in AI and Non-AI Sections

A similar pattern emerges in the distribution of time spent outside of class (Figure 11). The modal response for non-AI students was 4–6 hours per week (41%), whereas only 16% of AI students reported this level. Instead, AI students were much more likely to report 2–4 hours per week (44% vs. 23% in the non-AI section). Average reported time outside of class was also lower in the AI section (4.6 hours) than in the non-AI section (5.1 hours); the difference is significantly different at the 10% level ($p = 0.068$).

Finally, as it pertains attendance, students in the AI section reported a significantly larger level, as seen in the "All or almost all" category in Figure 12. This is in line with our prior finding that students in the AI class registered a superior performance in the daily activity grade (iclicker), which could only be obtained if the student was present in class.

Finally, Figure 12 shows notable differences in reported attendance between the two sections. In the AI class, 68% of students indicated attending "all or almost all" class sessions, compared to only 45% in the non-AI section. Conversely, a larger fraction of non-AI students reported attending only half or three-quarters of the sessions (41% vs. 24% in the AI class). This pattern suggests that students in the AI section were more consistently present in class. The finding is consistent with the earlier result that AI students earned higher scores on iClicker-based daily activities, which by design required in-class participation.

In sum, these results indicate that students in the AI section, while anticipating similar grade outcomes as their peers, reported higher satisfaction with the course, greater engage- ment (through higher participation and attendance), and lower self-reported effort and time commitments. These patterns are consistent with the interpretation that AI tools functioned as a substitute for some of the study time and labor typically required in a traditional setting.

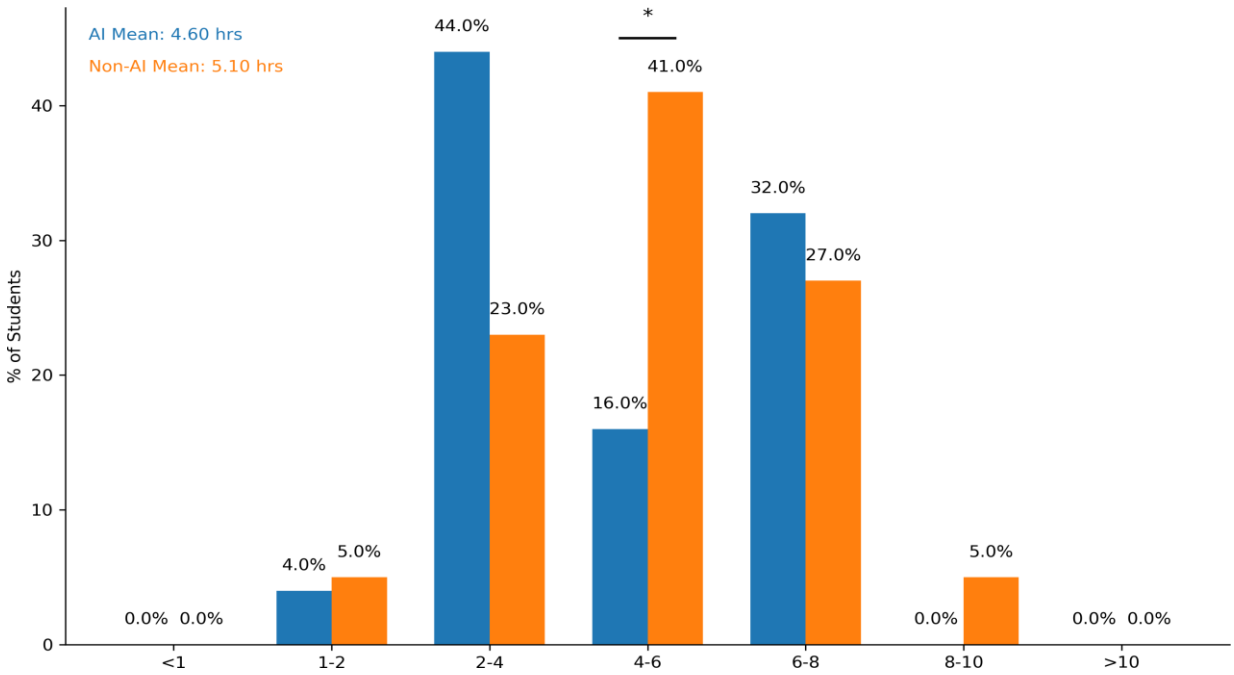


Figure 11. Time Spent Working on the Course Outside of Class

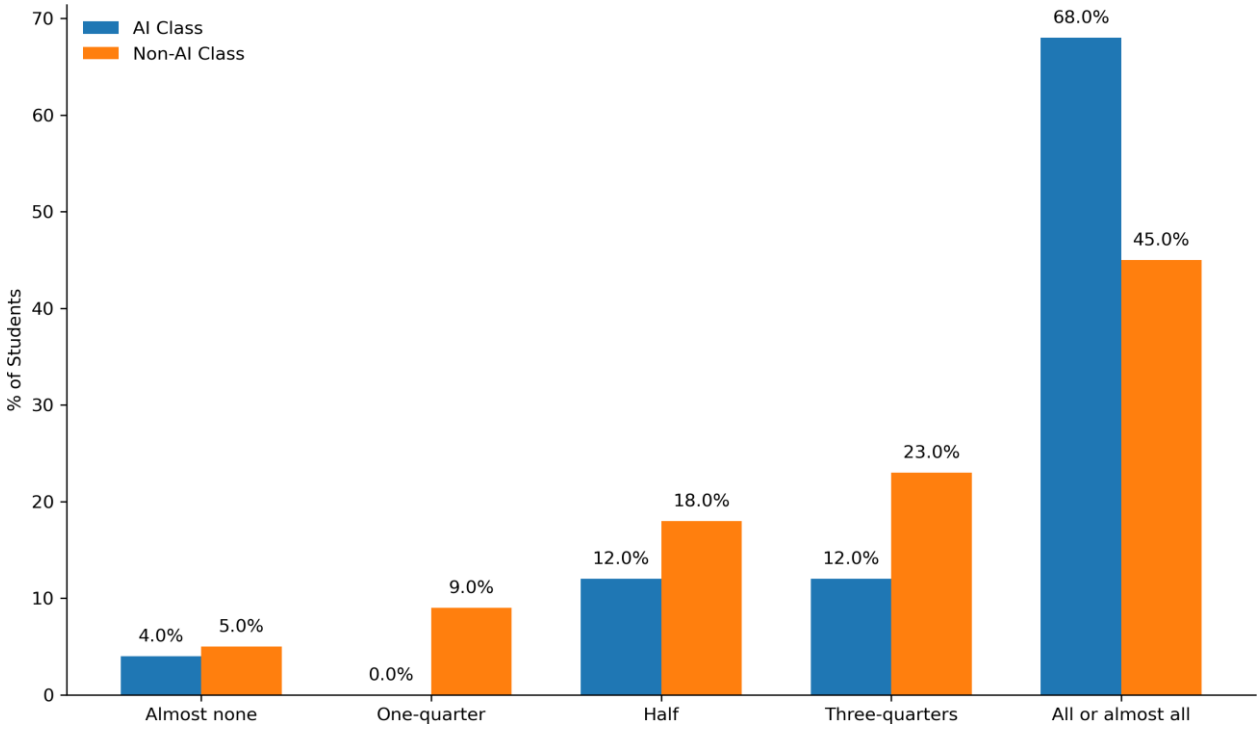


Figure 12. Reported Class Attendance in AI and Non-AI Sections

CONCLUSION

We set out to test whether structured access to generative AI improves *measured learning* in a traditional sense. The core result is simple: permitting and scaffolding AI use did *not* raise proctored exam scores in this course. Yet the intervention meaningfully changed *how* students learned and *how* they felt about their learning. Relative to the non–AI sec- tion, students with structured AI access concentrated their use into longer sessions, shifted toward revision–oriented uses (especially grammar/editing), reported more metacognitive behaviors (modifying outputs, catching errors, and—importantly—preferring their own an- swers), expressed more favorable views of AI's effects (particularly for efficiency, engagement, and confidence), and stated stronger intentions to keep using and studying AI—including a markedly higher inclination toward AI–intensive careers. In the classroom, the AI section participated more in real–time activities (iClicker) and reported higher satisfaction on the standardized university evaluation, while requiring no more—and in some cases fewer—hours outside class to achieve comparable grades.

Taken together, the evidence points to a productivity channel: AI helped students achieve "the same for less," primarily by reallocating effort and deepening the thought process rather than by boosting test performance. This distinction matters for how we evaluate educational technology. If AI shifts the learning production function—saving time, supporting revision, and encouraging scrutiny of model outputs—then focusing solely on exam scores misses a large part of the impact on the student experience.

To explain the observed gains on students' engagement in the AI section, we draw insights from Self-Determination Theory (SDT). The theory highlights autonomy, competence, and relatedness as three fundamental psychological factors that impact the levels of mental en- gagement in learning (Deci and Ryan, 1985; Ryan and Deci, 2000, 2017). We conjecture that our AI intervention could operate along all three dimensions. Providing structured access to AI expanded students' sense of autonomy, replacing a prohibitive baseline policy with the freedom to decide when and how to seek assistance. AI feedback and draft-level support likely enhanced students' competence, enabling them to build their work from clearer, more coherent starting points and to take greater ownership over subsequent revisions. And as students see the necessity to navigate a rapidly evolving AI landscape, the AI guided learn- ing may have felt more related to their future academic and career development. Hence, the engagement improvements we document are consistent with the mechanisms in SDT.

The findings also have implications for policy and pedagogy. Our implementation paired permission to use AI with clear guardrails (*proper disclosure/attribution*), explicit instruction (how to use AI for brainstorming, revision, and organizing information), and accountability (major exams were administered *without* AI). This bundle appears to cultivate more deliber- ate, self-regulated use and a more positive classroom climate, without sacrificing assessment integrity. For

instructors considering adoption, the practical takeaway is to *permit with scaf- folding* : be transparent about what counts as acceptable assistance and how to attribute it; require disclosure of AI contributions; model and assess critical and responsible engagement with AI outputs; retain proctored, no-AI assessments for high-stakes evaluation; and incor- porate reflective or audit tasks that ask students to critique, adapt, and justify AI-assisted work.

At the same time, several design limitations exist in this setting. The treatment was assigned at the section level for a single course and semester, with a modest sample ($n = 57$); our tests were therefore not powered to detect small causal effects (e.g., on exam grades), and we conducted multiple comparisons without family–wise adjustments. Survey outcomes are self–reported and may be sensitive to demand characteristics (e.g., students responding more favorably when AI is permitted), even though many items referenced "other courses" to mitigate this concern. The intervention is a bundle (permission + scaffolding + disclosure), so we cannot isolate the contribution of each component. Most importantly, the one–semester horizon prevents us from speaking to *long–run* consequences—especially for *critical thinking* and independent problem solving. If, as some recent work suggests, confidence in AI outputs can reshape cognitive processes away from information gathering and problem solving toward verification and output integration, then the durability and direction of those shifts remain unknown in our context.

These limitations suggest clear priorities for future work. Larger, multi–course random- ized trials—ideally across disciplines—should (i) pre–register primary outcomes spanning performance, process, and affect; (ii) include blinded or third–party grading of standardized writing/problem tasks; (iii) complement surveys with behavioral traces (e.g., platform logs) to measure actual use; and (iv) examine heterogeneity by prior achievement and student background. To speak directly to critical thinking and long–term impact, future studies should add pre/post measures of analytic reasoning and problem solving, calibrated trust- in-AI tasks, delayed retention and transfer assessments, and longitudinal follow–ups across subsequent courses. Factorial designs that decouple permission, scaffolding, and disclosure would isolate mechanism, and longer horizons could test whether early AI scaffolding builds durable skills (prompting, verification, revision) or crowds out independent problem solv- ing, and whether short–run efficiency gains translate into improved capstone, internship, or labor–market outcomes.

To sum up, in this experiment, structured AI access changed the *learning process*—improving efficiency, metacognitive engagement, and course experience—without moving proctored test scores. For instructors and institutions, the question is not only "Does AI raise exam perfor- mance?" but also "How should we teach, assess, and support students in an AI–rich learning environment?" Our results support transparent and guided adoption with assessments de- signed to reward critical use rather than irresponsible use.

REFERENCES

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, and Percy Liang. (2022). "On the Opportunities and Risks of Foundation Models." *arXiv preprint arXiv:2108.07258*.

Bond, Melissa, Hassan Khosravi, Maarten De Laat, Nina Bergdahl, Violeta Negrea, Emily Oxley, Phuong Pham, and Sin Wang Chong. (2024). "A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collabora- tion, and rigour." *International Journal of Educational Technology in Higher Education*, 21(1): 4.

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond. (2023). "Generative AI at Work." National Bureau of Economic Research NBER Working Paper 31161.

Cotton, David, Polly Cotton, and Richard Shipway. (2023). "Artificial Intelligence and Academic Integrity: The Ethics of Using AI in Student Assessment." *Assessment & Evaluation in Higher Education*, 48(1): 1–14.

Deci, Edward L., and Richard M. Ryan. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior*. New York, NY:Plenum Press.

Duin, Ann Hill, and Isabel Pedersen. (2023). "ChatGPT and Student Engagement: Motivational Benefits and Pedagogical Challenges." *Educational Technology Research and Development*, 71(4): 913–930.

Eaton, Sarah E., Katherine Crossman, and Rebecca Edino. (2023). "Academic in- tegrity in the age of artificial intelligence: Perspectives and challenges." *Journal of Aca- demic Ethics*, 21(2): 217–234.

Kasneci, Enkelejda, Katharina Sessler, Judith Kühner, Maria Bannert, and others. (2023). "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education." *Learning and Individual Differences*, 103: 102274.

Lee, Hao-Ping (Hank), Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. (2025). "The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers." *CHI '25*. New York, NY, USA:Association for Computing Machinery.

Mollick, Ethan, and Lilach Mollick. (2023). "Assigning AI: Seven Approaches for Stu- dents, with Prompts." *Journal of Management Education*, 47(3): 456–469.

Nguyen, Trang, and Mark A. McDaniel. (2023). "Integrating Generative AI Tools in Higher Education: Instructional Frameworks and Student Outcomes." *Educational Psy- chology Review*, 35(2): 543–562.

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

Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. (2023). “The Impact of AI on Developer Productivity: Evidence from GitHub Copilot.” *arXiv preprint arXiv:2302.06590*.

Ryan, Richard M., and Edward L. Deci. (2000). “Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being.” *American Psy- chologist*, 55(1): 68–78.

Ryan, Richard M., and Edward L. Deci. (2017). Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness. New York, NY:Guilford Press.

Selwyn, Neil, Luci Pangrazio, and Bronwyn Cumbo. (2023). “Passive Learners? Students’ Experiences and Perceptions of Generative AI in Education.” Learning, Media and Technology, 48(3): 333–347.

Vieriu, Ana-Maria, et al. (2025). “The Impact of Artificial Intelligence (AI) on Students’ Academic Development.” Education Sciences, 15(3): 343.

Zhai, Xiaojun, Pei-Chuan Lin, and Baochen Chen. (2023). “Teaching Critical Engage- ment with Generative AI Outputs.” Computers & Education, 202: 104803.