Generative AI's Impact on Student Achievement and Implications for Worker Productivity*

Naomi Hausman†, Oren Rigbi‡, and Sarit Weisburd§

Preliminary Draft – Work in Progress

March 2025

Abstract

Artificial intelligence (AI) is changing not only how we work, but also how we learn. Student use of AI in higher education is redefining the skills and knowledge of future workers. This study measures the effect of generative AI tools on human capital formation by analyzing university student performance. Student-course level data from a top Israeli university facilitates highly controlled performance comparisons—even within student—across courses more and less compatible with AI use for graded work, before and after the introduction of ChatGPT. AI leads to grade inflation and reduced grade variance, particularly benefiting lower-performing students. Grades' signal value to employers is diminished, complicating job matching. AI experience effects suggest both gains in AI-specific human capital and relative losses in non-AI human capital, as use of the tool may replace some learning. This work highlights costs AI may impose on future workforce productivity.

*We gratefully acknowledge outstanding research assistance from Nataliya Kris and funding from the Asper Center for Entrepreneurship.

†The Hebrew University Business School, Jerusalem, Israel, 9190501, naomi.hausman@mail.huji.ac.il

‡The Hebrew University Business School, Jerusalem, Israel, 9190501, oren.rigbi@mail.huji.ac.il

§The Hebrew University Business School, Jerusalem, Israel 9190501, sarit.weisburd@mail.huji.ac.il

1. Introduction

Artificial intelligence (AI) is changing not only how we work, but also how we learn (Bastani et al., 2024; Nie et al., 2024). Student use of AI in the course of their higher education is redefining the skills and knowledge of future workers. AI tools in education—such as chat bots, personalized learning platforms, automated tutoring, assisted programming, and content development—offer students new ways to both build technical skills and produce assignments. Early exposure to these tools could equip future workers with the skills to use AI efficiently, enhancing their workplace performance. But there is a growing concern that students may become overly dependent on AI, using it as a substitute for essential learning processes, such as problem-solving and critical thinking. This over-reliance could lead to knowledge gaps, ultimately lowering productivity and leaving future workers less prepared.[1] In addition, AI use on take-home exams and assignments could obfuscate underlying ability, reducing the signal value of grades and complicating job matching. The net effect on ultimate workforce productivity of these potential benefits and costs of AI use in universities is unknown.[2]

We study the effect of AI availability on human capital development and signaling in higher education via early adoption of generative AI tools by university students. We measure the extent to which student performance changes when students are presented with opportunities to employ AI in their graded work. Courses in which assignments and the final exam are completed at home are especially conducive to AI use ("AI-compatible courses"), while those with only in-person work compel students to rely on their human capital alone, without the help of AI ("AI-incompatible courses"). The difference in student performance (and grade distribution) between these two types of courses, after relative to before the introduction of AI tools, estimates the AI effect.

Detailed data on course syllabi and grades for all BA, MA, and MBA courses at one of Israel's largest universities facilitate tight identification with student and semester fixed effects and careful comparisons between similar courses. While baseline estimates control for unobserved heterogeneity across students and time, the many observable course characteristics in the data enable a propensity score matching strategy in which AI courses are compared to

[1] Bastani et al. (2024); Lehmann, Cornelius and Sting (2024); Darvishi et al. (2024) provide evidence on knowledge gaps resulting from AI use.

[2] Acemoglu (2024) estimates only modest TFP gains due to the complexities of AI use in certain contexts.

similar-on-observables courses that nevertheless used assessments not conducive to AI.[3] The data follow approximately 36,000 students in 6,000 courses from September 2018 to June 2024, spanning the introduction of ChatGPT for widespread use in November 2022. Because the data begin three years prior to ChatGPT's introduction, we can assess the parallel trends assumption underlying our estimation.

Since performance in AI-compatible courses reflects the combination of human capital and AI, a relative increase in performance may be measured even if AI partially or completely substitutes for non-AI related human capital development. To the extent that workforce tasks and university assignments are similarly amenable to AI use, a performance increase foretells corresponding increases in labor market productivity. However, university education tends to be grounded in critical thinking precisely to train students to perform in new situations; students who miss this training due to over-reliance on AI during their studies may be less able to translate their university performance into labor market productivity. We take two approaches to assessing the extent to which AI substitutes for human capital development. First, we measure student rank inconsistencies across more and less AI-compatible courses. Students whose within-course rank is much higher in AI-compatible courses than in others may be unprepared for occupations in which some tasks still require non-AI-related, basic human capital. These rank mismatches provide an estimate of knowledge gaps. Second, we estimate the effects of experience with AI in introductory courses on subsequent performance in both AI-compatible and AI-incompatible advanced courses. While the former is informative on development of AI-specific human capital, in which students learn to use the tool more productively, the latter is informative on the substitution of AI for basic human capital accumulation. Lower performance in AI-incompatible advanced courses, in which students must rely on their own knowledge and skills, suggests AI use in introductory courses likely crowds out some core learning.

We find a significant increase in student grades in both baseline and propensity score matching estimates of about 0.6-0.9 points in AI-compatible courses relative to others. This change is meaningful relative to the average grade of 86, potentially pushing students over the margin between a B and a B+, for example. Consistent with Brynjolfsson, Li and Raymond (2023); Noy and Zhang (2023); Cui et al. (2024), the increase is largest for students toward the bottom of the distribution, giving 25th percentile students more than 2.5 additional points in our most controlled specification. Similarly, we find a significant 33% decline in

[3]Observables include features of faculty (group of departments), degree program (BA/MA), credits, language of instruction, whether the course is mandatory, ex-ante enrollment, and course type (lecture/seminar).

the fraction of students who fail AI-compatible courses. We provide evidence that the grade distribution is not only pushed rightward but also compressed, pooling students of a variety of abilities under similar grades and reducing grades' signal value. This compression of the grade distribution has broader implications for the labor market, where the traditional role of grades in differentiating talent and predicting productivity may be diminished, potentially complicating hiring, forcing employers to find other means of assessment, and reducing the efficiency of worker-job matching.

Meanwhile, we provide evidence that experience with AI can have both positive and negative effects on future productivity, depending on whether future tasks permit the use of AI. While AI training unsurprisingly makes students better AI users, outperforming in advanced AI-compatible courses, it does seem to replace some basic human capital accumulation. Students who can use AI in introductory courses perform worse in advanced courses that are not compatible with AI use in graded work. Further supporting human capital replacement, we illustrate increasing rank inconsistencies for a given student across AI-compatible and AI-incompatible courses. When ChatGPT becomes available, a student's performance in AI-compatible courses becomes less predictive of his performance in AI-incompatible courses, in which he has to rely on his own skills.

Measuring the effect of generative AI tools on human capital development is a fundamentally difficult task for at least two reasons. First, students who choose to use AI for assignments are different than others in ways we can't observe or account for; they may use AI to compensate for lower ability, or they may be the most talented subset who have the skills and foresight to adopt AI most quickly and effectively. Naive estimates of the AI effect that don't account for selection into use may thus be upward or downward biased. Second, researchers generally don't observe how AI tools are used and thus to what extent they may complement or substitute for human capital in production of an output, in practice. When only the output is measured, it's then difficult to discern the extent to which it reflects human capital, AI, or the interaction of the two, and so inferring underlying human capital accumulation requires some empirical strategy. This paper uses granular data on student performance, along with naturally occurring variation in courses' compatibility with AI use, to identify the effect of AI. It further to isolates the effect on components of performance that reflect human capital alone. Since we still don't observe individual students' AI use, our estimates can be viewed as analogous to intent-to-treat (ITT) effects, where course compatibility with AI is a sort of instrument for use, but where we measure the direct effect of AI availability on performance. As a result, our estimates should be viewed as a lower bound on the effect of AI use per se

among students.[4] This approach eliminates selection bias, improving upon the confounded comparison of AI users to non-AI users, who are fundamentally different in many ways that also affect their performance.

The question of how generative AI tools affect university student performance and human capital acquisition is highly relevant to the design of universities' educational programming and assessment as well as to employers. To the extent grades become less useful signals of human capital because AI facilitates pooling, employers will have to develop alternative means of screening for both basic and AI-related human capital. Universities must balance the need to train students in modern technological tools that they'll use on the job and the potential for complementarities, on one hand, with the risk that under-monitored students facing outdated assignments may knowingly or unknowingly over-rely on these tools to substitute for important learning processes. While avoiding training students in AI augmentation would limit their AI-related human capital accumulation, allowing unfettered use would limit their traditional human capital development. Understanding the location of the status quo on this spectrum is a critical starting point for forming modified curricula and education policy. AI adoption among students in university today has implications for the productivity of the workforce tomorrow.

This proposal is organized as follows. Section 2 discusses relevant literature. Section 3 provides an overview of our empirical strategy and the grades database created for this project. Section 4 presents preliminary results on the impact of ChatGPT introduction on student performance and grade distributions (including robustness tests). Section 5 presents our plans for future work.

2. AI Tools in Education and the Labor Force

The growing literature on the adoption of generative AI tools has focused primarily on how they enhance the productivity of workers across industries, with a particular emphasis on tasks that align closely with the technology's strengths (e.g. Agrawal, Gans and Goldfarb (2019); Brynjolfsson, Li and Raymond (2023); Dell'Acqua et al. (2023); Demirci, Hannane and Zhu (2025); Acemoglu (2024)). Research on generative AI highlights its relatively fast adoption rate and heterogeneous effects across types of workers. Already in 2024, 39% of the US population between the ages of 18-64 use generative AI tools (Bick, Blandin and Deming

[4]The latter effect is analogous to a treatment-on-the-treated (TOT) effect, which is equal to the ITT scaled up by the probability of AI use given being in an AI-compatible course, under some assumptions.

(2024). Higher-ability firms and workers tend to benefit disproportionately from new technology adoption, in general (Acemoglu et al., 2023; Acemoglu, Lelarge and Restrepo, 2020; Koch, Manuylov and Smolka, 2021). For AI specifically, there is evidence in both directions, depending on context. (Otis et al., 2023) find that GPT-4 improves outcomes for high performing entrepreneurs in Kenya but worsens outcomes for low performers. But substantial research finds that AI tools have an equalizing effect, boosting worker productivity most among less skilled or experienced employees (Brynjolfsson, Li and Raymond, 2023; Noy and Zhang, 2023; Cui et al., 2024).

The positive impact of generative AI tools on productivity is not observed for all types of tasks. Dell'Acqua et al. (2023) find that AI tools improve both productivity and output quality for workers when applied to tasks well-suited to AI's capabilities. However, they also report worse performance outcomes on tasks requiring nuanced judgment, such as synthesizing insights from both structured and unstructured inputs. AI tools lead to more homogenized ideas, they show, reducing variability in participants' outputs. These complexities align with broader macroeconomic projections. Acemoglu (2024) estimates modest total factor productivity gains from AI—less than 0.53% over the next decade—due to diminishing returns on complex, context-dependent tasks. These findings suggest a nuanced role for AI in productivity enhancement and highlight the importance of understanding heterogeneity in adoption effects.

Even more complex is understanding the productivity effects of AI tools in learning environments. Productivity in an educational task, of course, is not just about the quality of the output but also, especially, about human capital gained in the process of producing the output. Studies of specialized AI interfaces that encourage learning through critical thinking rather than allowing solution-seeking report an increase in exam performance, suggesting human capital gains (Bastani et al., 2024; Nie et al., 2024). These findings contrast with others on the implications of general AI technology use by students. The latter indicate performance improvements on tasks for which AI tools are available but worse outcomes when the tool is no longer available, even relative to students who never had AI access (Bastani et al., 2024; Lehmann, Cornelius and Sting, 2024; Darvishi et al., 2024). Such studies support the concern that students substitute typical AI-tools for key learning processes.

We contribute to this literature by leveraging a large, comprehensive dataset that tracks student outcomes both before and after the introduction of ChatGPT across a range of courses with varying levels of suitability to generative AI tools. By analyzing grades on both

at-home exercises and in-class exams, we can disentangle the productivity benefit ChatGPT offers from concerns that it can serve as a crutch, creating knowledge gaps and leading to worse performance when the tool is absent. Additionally, the diverse student body at the University allows us to examine heterogeneity in AI's benefits across demographic and socioeconomic groups. Measuring distributional effects extends our understanding of whether ChatGPT exacerbates existing disparities in skill acquisition and performance or serves as an equalizing force in the educational context. Our analysis emphasizes the potential dampening effect of generative AI adoption on grade dispersion. Reduced variability in student performance could have significant implications for the quality of the match between graduates and employers, as homogeneity in grades may obscure distinctions in student ability and readiness for complex, real-world tasks. By addressing these gaps, our study aims to explore how generative AI tools will change the type of workers we employ and the tasks they will able to complete.

3. Empirical Strategy and Data

3.1. Empirical Strategy

An ideal experiment to measure the effect of student ChatGPT adoption on performance in university would be to randomize its use or availability across students or courses. While researchers can't control who uses ChatGPT, we can leverage differences across courses in the extent to which they are compatible with ChatGPT use in graded assignments with the introduction of ChatGPT on November 30th, 2022.

To identify the causal impact of ChatGPT adoption on student outcomes, we apply a dynamic Difference-in-Differences (DID) approach, capitalizing on variation across courses in the compatibility of assignments with ChatGPT use. Thus, we use a student level database and model final grade ($grade_{ict}$) for student i in course c, that she completed at time t using the following regression model:

$$\text{grade}_{ict} = \sum_{\substack{t=2019 \\ t\neq 2022}}^{2023} \delta_t \, \text{D}_t \times \text{AI_Course}_{ct} + \sum_{\substack{t=2019 \\ t\neq 2022}}^{2023} \gamma_t \, \text{D}_t + \beta \text{AI_Course}_{ct} + \theta' X_{ct} + \rho_i + \epsilon_{ict} \quad (1)$$

where $AI_Course_{ct} = 1$ if the course is AI compatible at time t and 0 otherwise. We categorize courses as AI-compatible if assignments predominantly involve tasks where generative AI could substantially assist (i.e., at-least 40% of work contributing to final grade is completed at home).[5] The binary variables D_t take the value of 1 when the course was completed in year t and 0 otherwise. The vector X_{ct} includes time varying characteristics of the course that may be correlated with grade outcomes such as number of students enrolled, fall or spring semester, department, language taught, and whether or not the course is mandatory.

The coefficient on the interaction term in the post-ChatGPT period, $\hat{\delta}_{2023}$, estimates the change in grades relative to the baseline year (2022) that was driven by the introduction of ChatGPT. We can interpret this as the causal effect of the availability of this new technology on student grades as long as we are able to control for general trends in grade outcomes occurring over time. Importantly, all students enrolled in an AI-compatible course ($AI_Course_{ct} = 1$) are treated, regardless of whether or not they personally use ChatGPT on their assignments. Thus, our analysis provides an intent-to-treat (ITT) effect of the availability of generative AI tools on grade outcomes.

The dynamic DID strategy employed in Equation 1 provides an unbiased estimate of the effect of the introduction of this new technological tool as long as, absent ChatGPT adoption, student performance trends in high and low compatibility courses would have evolved similarly. Ensuring that $\hat{\delta}_{2023}$ provides a causal estimate of the AI compatibility effect is complex because compatibility was not randomly assigned and we are unable to observe how high compatibility courses would have evolved absent ChatGPT. Importantly, the three years of data prior to the introduction of this tool allow us to test this hypothesis directly in the data by examining the interaction terms: $\hat{\delta}_{2019}$, $\hat{\delta}_{2020}$, and $\hat{\delta}_{2021}$. We would expect these interaction terms from the pre-period to not be significantly different from zero if the AI-compatible courses were following a similar grading trend to the non-compatible courses in the pre-ChatGPT period.

In the following section, we show that the parallel trends assumption holds and that we cannot reject the hypothesis that δ_{2019}, δ_{2020}, and δ_{2021} are equal to 0. To further improve the nature of the comparison between courses, we match treated courses ($AI_Course_{ct} = 1$) with control courses ($AI_Course_{ct} = 0$) based on the faculty where the course is provided (e.g. a treated course from the social sciences will be matched with a control course from the social sciences) and a course propensity score. The course propensity score is defined

[5] We find our results are robust to a range of alternative definitions of AI compatible courses.

as the probability of a course being classified as treated and is calculated as a function of: whether the course is mandatory for degree, whether it follows a standard lecture format (as opposed to a lab or seminar), class size, and language taught. We use this matching strategy to match each treated course to a control course with a similar propensity score and create a new variable $Match_Group$.[6] We then re-estimate Equation 1 when including $Match_Group$ fixed effects so that we are identifying the effect of this technological change on treated and control courses with very similar characteristics.

To examine whether the introduction of this new technology resulted in a change in the grade distribution , we use a course level panel database. We consider the following outcome variables $(y_{ct}) = (gradeB_{ct},\ gradeM_{ct},\ gradeT_{ct})$. These outcome variables measure the bottom quartile of grades for each course $(gradeB_{ct})$, the median grade for each course $(gradeM_{ct})$, and the top quartile of grades for each course $(gradeT_{ct})$. Thus, we estimate the following equation:

$$y_{ct} = \sum_{\substack{t=2019 \\ t\neq 2022}}^{2023} \alpha_t \, \mathrm{D}_t \times \mathrm{AI_Course}_{ct} + \sum_{\substack{t=2019 \\ t\neq 2022}}^{2023} \rho_t \, \mathrm{D}_t + \omega \mathrm{AI_Course}_{ct} + \Omega' X_{ct} + \epsilon_{ct} \tag{2}$$

This equation provides an opportunity to estimate changes in the variance of grades in courses that are more AI compatible in comparison to those that were less affected by the introduction of this new tool. Using this course level dataset we follow a similar identification strategy as when estimating Equation 1 to ensure causal results. Specifically, testing the parallel trends assumption regarding grade variance across these courses prior to the introduction of ChatGPT, and analyzing results when including $MatchGroup$ fixed effects.

3.2. Data Description

Our analysis leverages detailed data obtained from a large research university in Israel, covering five academic years from 2018-2019 to 2022-2023. Notably, the academic year in Israel begins around October 20th, meaning that students had access to ChatGPT for the majority of the 2022-2023 academic year, except during its first month. This dataset enables a comprehensive analysis of the impact of generative AI tools on students outcomes by integrating information at the student and course levels.

[6] We use a nearest neighbor matching strategy with replacement. In situations where a treated course is matched to multiple control courses, we randomly choose one control course to be in the $Match_Group$.

At the **student level**, the dataset includes detailed records of all courses taken, including the semester of completion, grades from each attempt at the course's final exam, and the final course grade, which may include other grade components. Students are typically entitled to take the final exam twice, and the highest grade from these attempts is used in the calculation of the final grade; our dataset captures all exam attempts. Additional student background characteristics include gender, birth year, primary department of study, and indicators for specific demographic groups: Arab, Bedouin, Druze, Ethiopian, immigrant (including the year of immigration), and ultra-Orthodox high school graduates.

At the **course level**, the dataset incorporates syllabus-derived details such as the weight of each grade component in determining the final grade, the course format (e.g. lecture, lecture and section, lab), and whether the course was mandatory. It also includes information on the faculty offering the course as well as the name of the professor teaching the course. Lastly, it provides a detailed split of the components contributing to the final grade and assigns a specific weight to each component. Figure 1 presents a histogram of the cumulative share of the course grades derived from non-AI compatible components (in-class exams, midterms, and lab work). This distribution supports the definition of treatment and control groups: course with $<= 60\%$ of grades from non-AI compatible components are categorized as treatment courses, while those with $>= 90\%$ fall into the control group. These cutoffs allow us to include a substantial proportion of courses in the analysis.

?? provides summary statistics of key course characteristics, comparing treatment and control groups. While some differences are observed in course characteristics between the groups, the background characteristics of students enrolled in these courses are largely similar. This comparability minimizes the likelihood of confounding effects due to pre-existing student differences, strengthening the validity of our analysis.

4. Results

In this section, we present preliminary results estimating the impact of generative AI tools on grade outcomes. Figure 4 summarizes the event-study results from estimating Equation 1. We utilize our full database of treatment and control courses, encompassing 524,065 student-course observations. In our treatment courses where assignments align closely with ChatGPTs capabilities, we estimate a 0.7 grade point increase (s.e. 0.27) after the introduction of ChatGPT. This effect is statistically significant at the 1 percent level.

To interpret the grade increase estimated from Equation 1 as a direct result of the availability of ChatGPT, it is necessary to ensure that the control courses provide an accurate baseline for expected grade trends absent AI tools. Because our analysis is not the result of a randomized experiment, there are likely to be differences between courses that are defined as AI-compatible ($AI_Course_{ct} = 1$) versus courses that are not ($AI_Course_{ct} = 0$). Indeed, in **??** we observe that AI-compatible courses (Treatment) are significantly smaller and are more likely to be electives. Despite this difference, Figure 4 shows that in the years 2019, 2020, and 2021 the difference in grades between AI-compatible and non-AI-compatible courses was not statistically significant from the difference observed in 2022. This suggests that the parallel trends assumption that is necessary for a difference-in-difference analysis holds in this environment, supporting the argument that our results reflect a response to the availability of generative AI tools.

To strengthen the validity of our findings, we re-estimate Equation 1 using $Match_Group$ fixed effects as defined in section 3. The results, shown in Table 2 demonstrate the robustness of our estimates across a range of specifications. In column (1) of Table 2 we provide the estimates for the event study specification appearing in Figure 4 without $Match_Group$ fixed effects. In columns (2) through (5), we summarize our results when including $Match_Group$ fixed effects and incorporating additional controls such as the estimated propensity score for the course and other course characteristics. The estimated impact of ChatGPT introduction across these specifications remains significant at the 10 percent level, and ranges between 0.6 and 0.8 points.

To investigate how the introduction of generative AI tools impacts the distribution of grades within courses, we re-run the analysis at the course level, focusing on changes in the bottom quartile, top quartile, and median grade (see Equation 2). **??** demonstrates that the grade increases observed in Table 2 are not evenly distributed across students. The bottom 25th percentile exhibits a 0.85 grade increase, which is significant at the 1 percent level, while the top 25th percentile does not exhibit any statistically significant grade change. This suggests that lower-performing students benefit more from ChatGPT's introduction than their higher-performing peers.

This heterogeneity across students becomes more pronounced when applying $Match_Group$ fixed effects. In column (4) of Table 2, applying our most conservative specification of comparing treated courses to control courses with similar propensity scores, we find a 2.3-point grade increase for students in the bottom quartile with no statistically significant effect

observed at either the grade median or top quartile of grades. This finding that generative AI tools primarily aid lower-performing students, suggests that it could play a role in evening the playing field by improving human capital accumulation among the weakest students. However, this compression of the grade distribution raises concerns regarding the signaling value of grades to employers. This problem is even more pronounced if the grade increase estimated in the our analysis does not reflect an increase in learning and skills, but rather, over-reliance on a tool that is only helpful when completing very specific types of tasks.

4.1. Heterogeneity

Our large database of students from different backgrounds taking courses from a range of fields provides an opportunity to examine heterogeneous responses to the public availability of generative AI tools.

Previous research has found that men are more likely to adopt ChatGPT than women (Nie et al., 2024; Bick, Blandin and Deming, 2024). However, our analysis focuses on the broader implications of using this technology, capturing both its adoption and effective usage. In Figure 5, we estimate Equation 1 separately for male and female students. Notably, we do not find a significant difference between genders in the impact of ChatGPT availability on grade outcomes. Given that one might expected a lower adoption rate among female students, this could suggest that women use generative AI tools more effectively.

In Figure 6, we take a closer look at how ChatGPT impacted the grade distribution. Thus, we provide results from student-level analysis examining a set of additional outcome variables: $grade55_{ct}$, $grade65_{ct}$, $grade75_{ct}$, $grade85_{ct}$, $grade95_{ct}$, and $grade100_{ct}$. Each of these variable receives a value of 1, when the student's grade in class c in year t falls within the given range and 0 otherwise. These alternative outcome variables provide an opportunity to observe whether the introduction of generative AI tools impacts a specific part of the grade distribution. Indeed, in-line with our analysis at the course level (**??**), our results suggests that the introduction of generative AI tools primarily helps the weakest students who otherwise would have failed the class. Thus, we observe a significant 0.1 percentage point decrease in the probability that a student received a grade below 56 in a treated course after the introduction of generative AI tools (at a baseline failure rate of 2.7%). This implies that the introduction of ChatGPT decreased the probability of a student failing an AI-compatible course by 37%. We also find a significant 16% decrease in the probability of receiving a

passing grade below 65. Our analysis suggests that there may be an increase in the percent of students earning grades above 86, but these estimates are very noisily measured.

5. Discussion

References

Acemoglu, Daron. 2024. "The simple macroeconomics of AI*." *Economic Policy*, eiae042.

Acemoglu, Daron, Claire Lelarge, and Pascual Restrepo. 2020. "Competing with robots: Firm-level evidence from France." Vol. 110, 383–388, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Acemoglu, Daron, Gary Anderson, David Beede, Catherine Buffington, Eric Childress, Emin Dinlersoz, Lucia Foster, Nathan Goldschlag, John Haltiwanger, Zachary Kroff, et al. 2023. "Advanced Technology Adoption: Selection or Causal Effects?" Vol. 113, 210–214, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Agrawal, Ajay, Joshua S Gans, and Avi Goldfarb. 2019. "Artificial intelligence: the ambiguous labor market impact of automating prediction." *Journal of Economic Perspectives*, 33(2): 31–50.

Bastani, Hamsa, Osbert Bastani, Alp Sungu, Haosen Ge, Ozge Kabakcı, and Rei Mariman. 2024. "Generative AI can harm learning." *Available at SSRN*, 4895486.

Bick, Alexander, Adam Blandin, and David J Deming. 2024. "The rapid adoption of generative AI." National Bureau of Economic Research.

Brynjolfsson, Erik, Danielle Li, and Lindsey R Raymond. 2023. "Generative AI at work." National Bureau of Economic Research.

Cui, Zheyuan Kevin, Mert Demirer, Sonia Jaffe, Leon Musolff, Sida Peng, and Tobias Salz. 2024. "The effects of generative AI on high skilled work: Evidence from three field experiments with software developers." *Available at SSRN 4945566*.

Darvishi, Ali, Hassan Khosravi, Shazia Sadiq, Dragan Gašević, and George Siemens. 2024. "Impact of AI assistance on student agency." *Computers & Education*, 210: 104967.

Dell'Acqua, Fabrizio, Edward McFowland III, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani. 2023. "Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality." *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, , (24-013).

Demirci, Ozge, Jonas Hannane, and Xinrong Zhu. 2025. "Who is AI replacing? the impact of generative ai on online freelancing platforms." *Management Science*.

Koch, Michael, Ilya Manuylov, and Marcel Smolka. 2021. "Robots and firms." *The Economic Journal*, 131(638): 2553–2584.

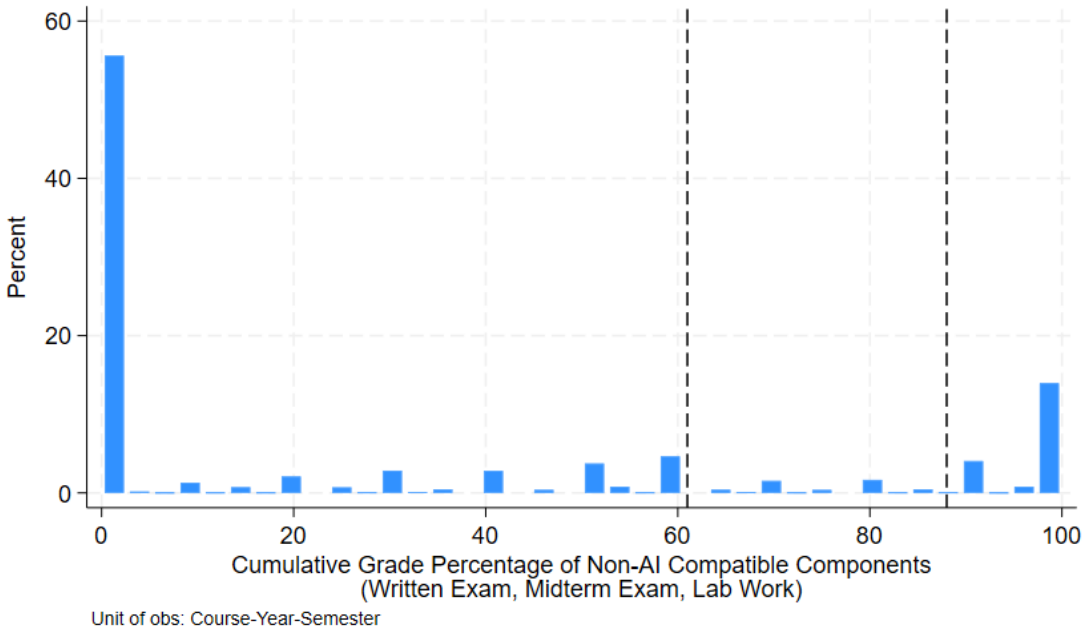
Lehmann, Matthias, Philipp B Cornelius, and Fabian J Sting. 2024. "AI Meets the Classroom: When Does ChatGPT Harm Learning?" *arXiv preprint arXiv:2409.09047*.

Nie, Allen, Yash Chandak, Miroslav Suzara, Ali Malik, Juliette Woodrow, Matt Peng, Mehran Sahami, Emma Brunskill, and Chris Piech. 2024. "The GPT Surprise: Offering Large Language Model Chat in a Massive Coding Class Reduced Engagement but Increased Adopters' Exam Performances." Center for Open Science.

Noy, Shakked, and Whitney Zhang. 2023. "Experimental evidence on the productivity effects of generative artificial intelligence." *Science*, 381(6654): 187–192.

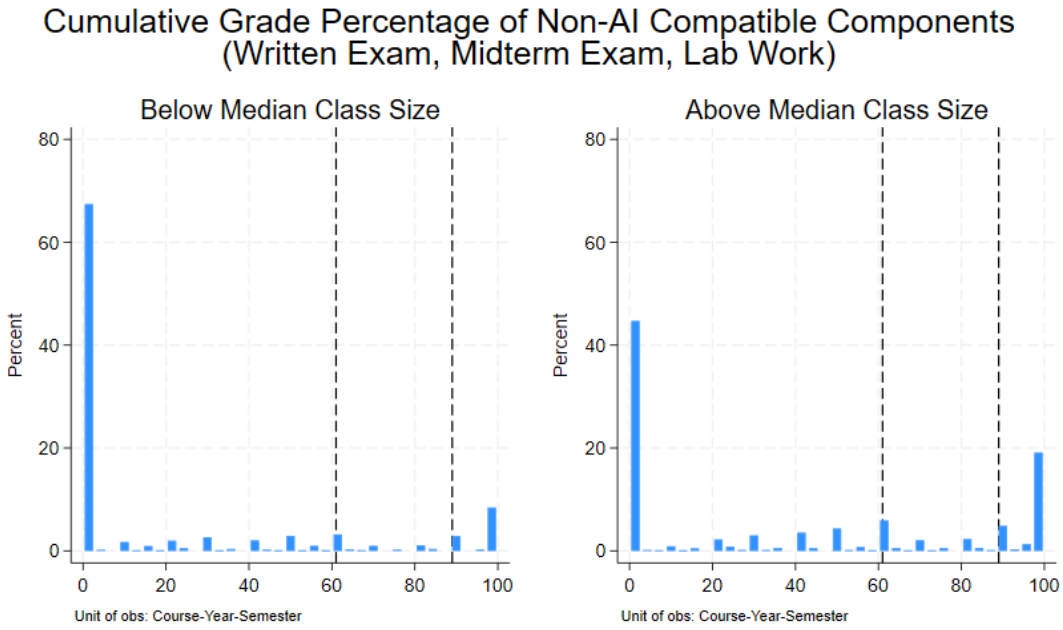
Otis, Nicholas, Rowan P Clarke, Solene Delecourt, David Holtz, and Rembrand Koning. 2023. "The uneven impact of generative AI on entrepreneurial performance." *Available at SSRN 4671369*.

Figure 1: Definition of Treatment and Control Groups Based on Grade Components



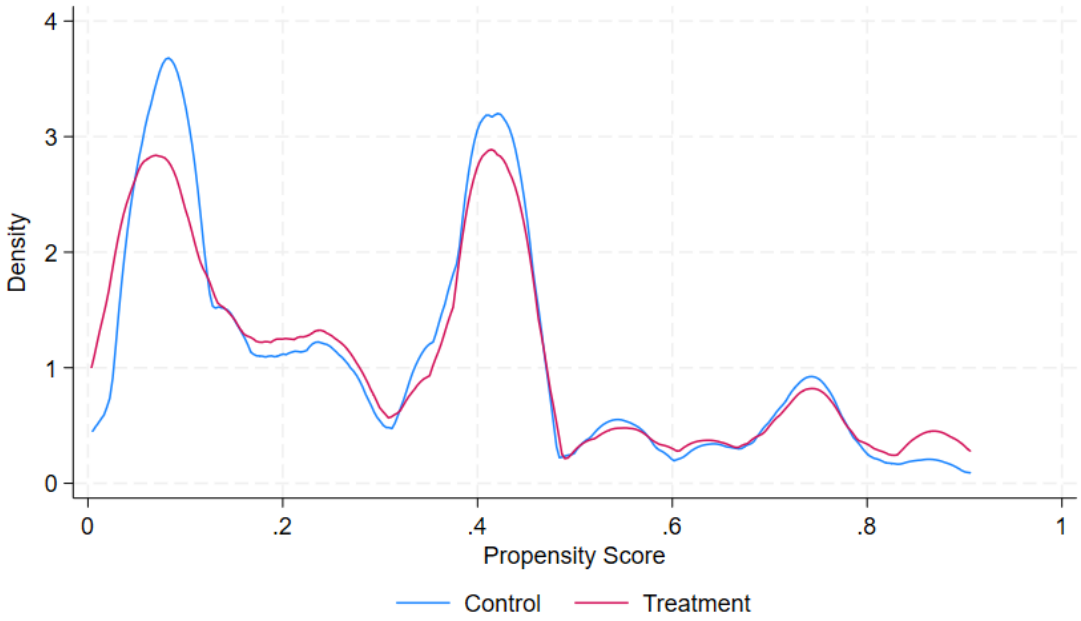
Notes: The histogram presents the distribution of courses according to the AI-incompatibility of their grade components. Courses in which all of the work is AI-incompatible are in the mass at 100. Courses in which all of the graded work is AI-compatible are in the mass at 0. The two vertical, dashed black lines represent cutoffs for treatment and control groups. Courses to the left of the first dotted line (60% AI-incompatible) are in the treatment group, and courses to the right of the second dotted line (90% AI-incompatible) are in the control group. Courses between the two dotted lines are omitted from the analysis. Grade components incompatible with AI usage are in-class exams and lab work. Courses with a cumulative weight of non-AI compatible components below 60% are classified as treatment courses, while courses with a weight greater than 90% are classified as control courses.

Figure 2: Distribution of Grade Components by Class Size



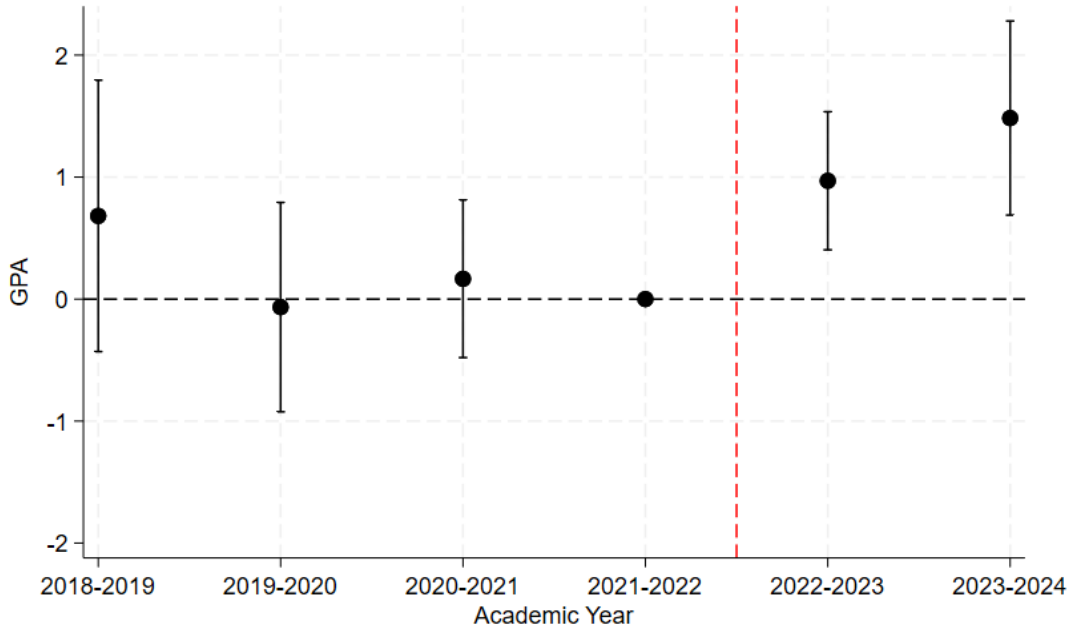
Notes: The histogram presents the distribution of courses according to the AI-incompatibility of their grade components in small versus large courses. Courses in which all of the work is AI-incompatible are in the mass at 100. Courses in which all of the graded work is AI-compatible are in the mass at 0. The two vertical, dashed black lines represent cutoffs for treatment and control groups. Courses to the left of the first dotted line (60% AI-incompatible) are in the treatment group, and courses to the right of the second dotted line (90% AI-incompatible) are in the control group. Courses between the two dotted lines are omitted from the analysis. Grade components incompatible with AI usage are in-class exams and lab work. Courses with a cumulative weight of non-AI compatible components below 60% are classified as treatment courses, while courses with a weight greater than 90% are classified as control courses.

Figure 3: Distribution of Propensity Scores in Matched Sample



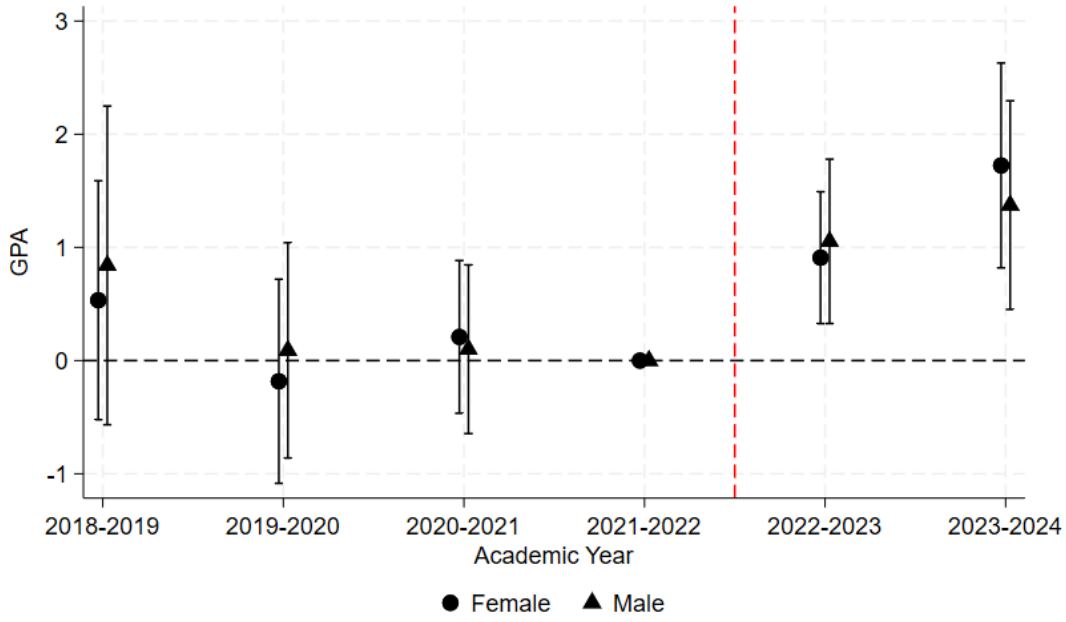
Notes: Figure shows the distributions of propensity scores for treated and control units in the matched sample.

Figure 4: Event Study: Effect of AI Availability on Grades



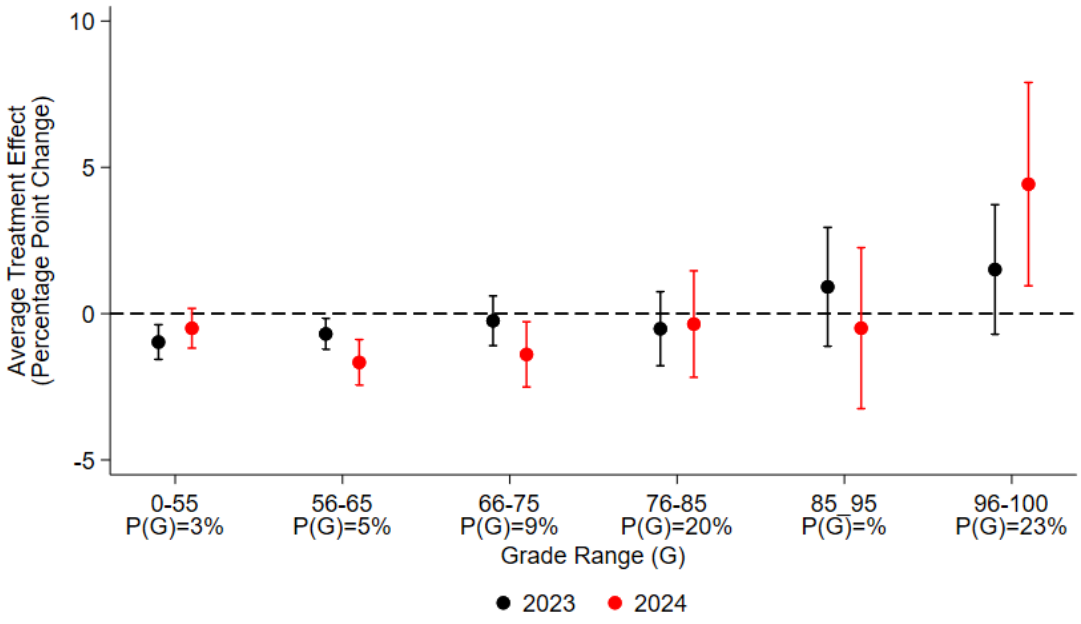
Notes: Points on the graph represent coefficients on $AI - compatible * year$ interaction terms in Equation 1. Error bars reflect 95% confidence intervals. The vertical dashed line marks the beginning of the widespread availability of ChatGPT.

Figure 5: Event Study: Effect of AI Availability on Grades, by Gender



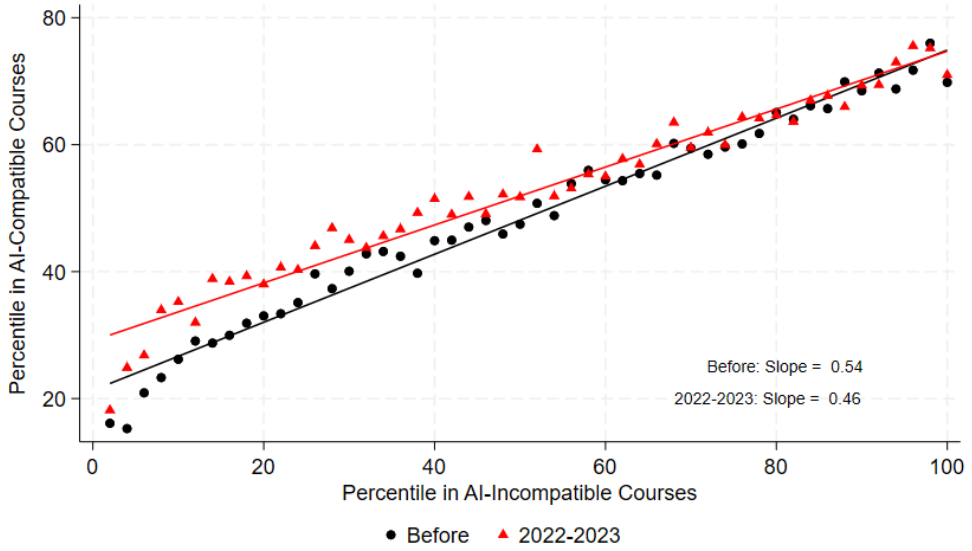
Notes: Points on the graph represent coefficients on $AI - compatible * year$ interaction terms in Equation 1 when run separately for male and female students. Error bars reflect 95% confidence intervals. The vertical dashed line marks the beginning of the widespread availability of ChatGPT.

Figure 6: Grade Distribution Effects: Probability of Scoring in each Grade Range

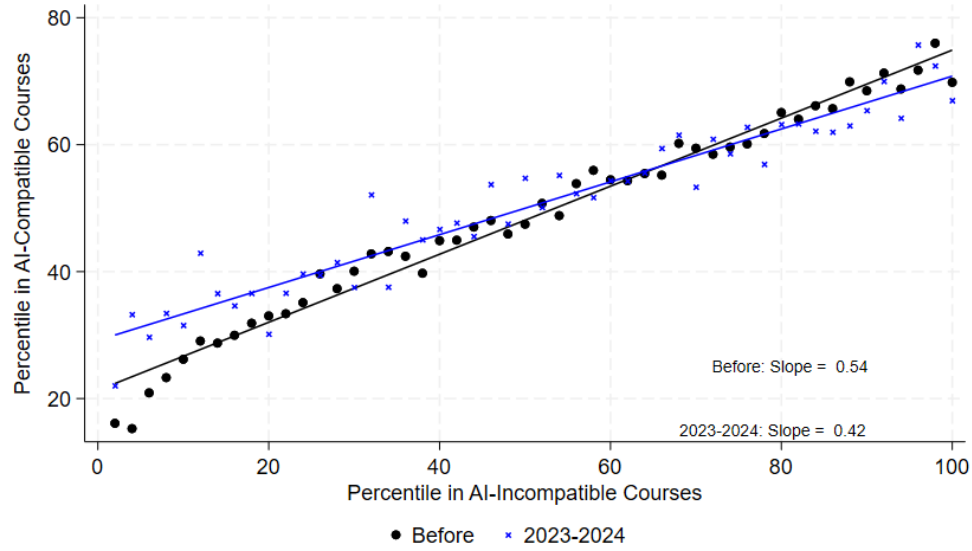


Notes: The figure presents coefficients on $AI - compatible * year$ interactions from a version of Equation 1 in which the dependent variable is the probability of scoring in the given grade range. Error bars reflect 95% confidence intervals.

Figure 7: Within-Student Rank Consistency in AI-Compatible vs. AI-Incompatible Courses



(a) Before ChatGPT vs 1st Year After (2023)



(b) Before ChatGPT vs 2nd Year After (2024)

Notes: The figure plots within-student rank-rank relationships for student grade rankings across AI-compatible (y-axis) and AI-incompatible (x-axis) courses. Flatter slopes indicate less predictability of rank in one course type for rank in the other.

Table 1: Mean Treated and Control Course Characteristics in Full and Matched Samples

Variable	**Control** (AI-incompatible)		**Treatment** (AI-compatible)	
	Full	**Matched**	**Full**	**Matched**
Avg. Grade	84	55	90	90
Class Size	68	55	31	32
Mandatory Class (%)	64	59	45	62
Age	27	27	29	27
Male (%)	44	47	42	40
Humanities (%)	10	17	22	23
Natural Sciences (%)	23	7	15	11
Social Sciences (%)	11	26	22	24
Agriculture (%)	33	13	13	12
Law (%)	6	4	5	5
Social Work (%)	6	11	9	9
Engineering (%)	1	18	6	10
Course*Semester	2,714	897	4,224	3,247
Student*Course*Semester	184,597	131,615	130,739	105,418

Table 2: Effect of AI Availability on Student Grades

	Dependent Variable: Grade (0-100)				
	Naive		Propensity Score Matching		
	(1)	(2)	(3)	(4)	(5)
Before:					
2018-2019*AI-compatible	0.682	0.375	-0.286	0.135	0.117
	(0.567)	(0.858)	(0.850)	(0.812)	(0.816)
2019-2020*AI-compatible	-0.066	0.530	0.296	-0.023	-0.048
	(0.438)	(0.634)	(0.557)	(0.563)	(0.565)
2020-2021*AI-compatible	0.166	0.256	0.077	-0.162	-0.168
	(0.330)	(0.490)	(0.527)	(0.472)	(0.473)
After:					
2022-2023*AI-compatible	0.970***	0.942**	0.809**	0.645*	0.633*
	(0.289)	(0.387)	(0.365)	(0.345)	(0.344)
2023-2024*AI-compatible	1.484***	1.334**	0.745	0.566	0.532
	(0.406)	(0.540)	(0.545)	(0.486)	(0.483)
Course Char. Controls	✓	✓	✗	✓	✓
Match Group FE	✗	✗	✓	✓	✓
P-Score Control	✗	✗	✗	✗	✓
R^2	0.477	0.554	0.565	0.581	0.581
N	500,611	361,591	361,591	361,591	361,591

Note: An observation in the sample is a student-course-semester. Column (1) uses the full sample, while columns (2)-(5) use the reduced sample of courses that were matched with propensity score matching. The dependent variable is the student's course grade, from 0-100. The baseline (omitted) interaction is 2021-2022*AI-compatible. AI-compatible (treated) courses are defined as those in which in-class exams and lab work comprise less than 60% of the grade, such that AI can be used for much of the graded work. AI-incompatible (control) courses are defined as those in which in-class exams and lab work comprise more than 90% of the grade. Propensity score matching is NN-1, and only courses with constant treatment status over time are included. Course characteristic controls include fixed effects for class size quartile, department, number of credits, class type, mandatory class, language of instruction, and degree (BA/MA). All specifications include student FE and semester FE. Standard errors are clustered by course. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3: Effect of AI Availability on the Grade Distribution

	Naive			Propensity Score Matching		
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable: Grade at:	p25	p50	p75	p25	p50	p75
Before:						
2018-2019*AI-compatible	0.875	-0.157	-0.694	1.093	-0.083	-0.569
	(1.077)	(0.784)	(0.678)	(0.993)	(0.798)	(0.674)
2019-2020*AI-compatible	0.081	-0.443	-0.594	-0.253	-0.720	-0.646
	(1.114)	(0.788)	(0.705)	(1.182)	(0.866)	(0.745)
2020-2021*AI-compatible	1.620	0.270	-0.507	1.124	0.014	-0.498
	(1.206)	(0.893)	(0.523)	(1.293)	(0.957)	(0.537)
After:						
2022-2023*AI-compatible	3.014***	1.272**	0.738	2.618**	1.044	0.719
	(0.933)	(0.612)	(0.453)	(1.037)	(0.675)	(0.495)
2023-2024*AI-compatible	2.383**	1.125	0.583	0.917	0.184	0.207
	(0.978)	(0.752)	(0.511)	(0.877)	(0.686)	(0.519)
Dep. Var. Avg.	84.72	90.18	94.14	84.72	90.18	94.14
R^2	0.426	0.401	0.312	0.588	0.582	0.508
N	10,076	10,076	10,076	10,076	10,076	10,076

Note: An observation in the sample is a course-semester. All columns are estimated on the reduced sample of courses that were matched with propensity score matching. The dependent variable is the grade at the percentile indicated at the top of each column. The baseline (omitted) interaction is 2021-2022*AI-compatible. AI-compatible (treated) courses are defined as those in which in-class exams and lab work comprise less than 60% of the grade, such that AI can be used for much of the graded work. AI-incompatible (control) courses are defined as those in which in-class exams and lab work comprise more than 90% of the grade. Propensity score matching is NN-1, and only courses with constant treatment status over time are included. Course characteristic controls in all specifications include fixed effects for class size quartile, department, number of credits, class type, mandatory class, language of instruction, and degree (BA/MA). All specifications include student FE and semester FE. Standard errors are clustered by course. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Heterogeneity in AI Effect on Grades

Dimension		No	Yes
Male	AI Compatible X 2022-2023	0.909***	1.053***
		(0.297)	(0.370)
	AI Compatible X 2023-2024	1.724***	1.374***
		(0.462)	(0.470)
	Observations	276,397	200,672
Young (< Age 26)	AI Compatible X 2022-2023	0.933***	1.132***
		(0.306)	(0.358)
	AI Compatible X 2023-2024	1.287***	2.079***
		(0.398)	(0.537)
	Observations	237,658	238,466
Advanced Course	AI Compatible X 2022-2023	0.999	0.843**
		(0.742)	(0.355)
	AI Compatible X 2023-2024	0.935	1.466***
		(0.876)	(0.511)
	Observations	144,254	254,662
STEM	AI Compatible X 2022-2023	0.769***	1.332
		(0.287)	(0.920)
	AI Compatible X 2023-2024	1.511***	1.269
		(0.409)	(0.969)
	Observations	373,967	119,260
Business	AI Compatible X 2022-2023	0.769**	0.188
(vs. non-Business non-STEM)		(0.298)	(1.101)
	AI Compatible X 2023-2024	1.398***	0.522
		(0.416)	(1.328)
	Observations	343,117	29,930
Senior Professor	AI Compatible X 2022-2023	0.791**	1.093**
		(0.392)	(0.486)
	AI Compatible X 2023-2024	1.613***	1.105*
		(0.578)	(0.576)
	Observations	272,268	212,846
Large Course(>25 students)	AI Compatible X 2022-2023	1.123**	1.075***
		(0.565)	(0.327)
	AI Compatible X 2023-2024	2.423***	1.671***
		(0.813)	(0.453)
	Observations	56,609	435,675

Note: Each cell in the table represents a separate regression illustrating heterogeneity in the treatment effects along the dimension indicated on the left. An observation in each sample is a student-course-semester. The dependent variable is the student's course grade, from 0-100. The baseline (omitted) interaction is 2021-2022*AI-compatible. AI-compatible (treated) courses are defined as those in which in-class exams and lab work comprise less than 60% of the grade, such that AI can be used for much of the graded work. AI-incompatible (control) courses are defined as those in which in-class exams and lab work comprise more than 90% of the grade. Course characteristic controls include fixed effects for class size quartile, department, number of credits, class type, mandatory class, language of instruction, and degree (BA/MA). All specifications include student FE and semester FE. Standard errors are clustered by course. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 5: The AI Experience Effect on Advanced Course Grades

	AI-Compatible Advanced Course Grade			AI- Incompatible Advanced Course Grade		
	(1)	(2)	(3)	(4)	(5)	(6)
Second Year*AI Experience	1.216***	0.676***	0.912***	-9.111***	-0.020	-0.439
	(0.267)	(0.243)	(0.229)	(0.328)	(0.365)	(0.348)
Course Char. Controls	✗	✓	✓	✗	✓	✓
Student FE	✗	✗	✓	✗	✗	✓
R^2	0.028	0.228	0.559	0.078	0.236	0.558
N	23,760	23,760	23,760	39,915	39,915	39,915

Note: An observation in the sample is a student-course-semester. Estimates presented are coefficients on the interaction of Second Year*AI Experience as described in Equation **??**. The dependent variable is the student's course grade, from 0-100, in the type of advanced course indicated. The sample is restricted to the years 2022-2024. In columns (1)-(3), the sample includes AI-compatible courses only from students' first and second year of studies. In columns (4)-(6), the sample includes AI-compatible courses from students' first year of studies and AI-incompatible courses from their second year. The "AI Experience" (treatment) group is defined as students who began their studies in 2022-23, once ChatGPT was available, and took at least one AI-compatible introductory course in that year. The control group is defined as students who began their studies in 2021-22, before ChatGPT was available, and took at least one AI-compatible introductory course in that year. Course characteristic controls include fixed effects for class size quartile, department, number of credits, class type, mandatory class, language of instruction, and degree (BA/MA). All specifications include semester FE. Standard errors are clustered by course. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$