UC Berkeley Center for Studies in Higher Education

Higher Education Working Paper

Volume No. 26-3

Artificial Intelligence and Grade Inflation

Igor Chirikov

Suggested Citation: Chirikov, I (2026). Artificial Intelligence and Grade Inflation. *CSHE Higher Education Working Paper Series*, Vol. 26-3. Retrieved from: escholarship.org/uc/item/80x8d3qd

Artificial Intelligence and Grade Inflation

Igor Chirikov*

May 2026

Abstract

Generative AI tools can undermine the informational value of grades by performing graded course tasks. I analyze the impact of AI on grade distributions across more than 500,000 grades at a large research university from 2018 to 2025 using a difference-in-differences design. Courses with more AI-exposed tasks, such as writing and coding, saw substantial grade increases after ChatGPT's release: the share of A grades rose by 13 percentage points, or about 30% relative to the 2022 baseline. These increases were larger where homework carried greater weight, consistent with AI substituting for student work rather than broad learning gains from AI.

JEL Codes: I23, I24, J24, O33

---

* Center for Studies in Higher Education, Goldman School of Public Policy, UC Berkeley (email: chirikov@berkeley.edu). I am grateful to Jesse Rothstein, Annaliese Paulson, and Samuel Ayers for their helpful comments. I also thank Lynn Chien, Christopher Mach, Sophie McKenna, and Smrithi Senthilnathan for their exceptional research assistance throughout this project, supported by UC Berkeley's Data Discovery Program. All errors are my own.

1 Introduction

Grades are a central element of skill certification in higher education. They summarize student performance and provide signals used by students, instructors, graduate programs, and employers. These signals are consequential: grades and grading standards impact students' academic choices, human capital allocation, and long-run earnings (Bleemer and Mehta 2024; Nordin et al. 2019; Denning et al. 2026). Generative artificial intelligence (AI) tools now threaten this certification function by performing many of the graded tasks universities use to assess skills (Geerling et al. 2023; Borges et al. 2024). If grades increasingly reflect AI-assisted output rather than students' own skills, their informational value declines, weakening the efficiency of human capital allocation.

Prior research identified a number of mechanisms that distort grade signals: instructor incentives tied to student evaluations and tenure (Langbein 2008; Carrell and West 2010; Keng 2018), signaling incentives across schools (Chan et al. 2007), institutional grading and disclosure policies (Bar et al. 2012; Butcher et al. 2014), and measurement error, discretion, and bias (Hanna and Linden 2012; Diamond and Persson 2016). These mechanisms operate through instructor behavior, institutional rules, or the measurement process itself. Generative AI potentially introduces a distinct technological shock to grading: it changes the production of graded work before instructors observe and assess it. Even with unchanged grading standards, grades can become inflated when AI improves submitted work on assessed tasks without a corresponding increase in students' underlying skills. This study explores whether this AI-driven channel has become a new mechanism of grade inflation and grade-signal erosion.

To study this channel, I adapt task-based models of technological change (Autor et al. 2003; Acemoglu and Restrepo 2019) from production to education. Courses can be viewed as bundles of learning tasks through which students develop skills. AI can affect these tasks through displacement, where it performs work instead of students; augmentation, where it supports task practice without replacing essential effort; and reinstatement of new AI-based tasks (Chirikov 2026). These channels have different implications for grades: displacement can raise grades without improving underlying skills, while augmentation and reinstatement may raise both performance and skill.

I study grade changes in over 500,000 student-course enrollments across a balanced panel of 319 courses spanning 84 departments at a large selective research university in Texas from 2018

to 2025. I measure AI exposure using task composition extracted from each course's Fall 2022 syllabus (published in August 2022 before ChatGPT's release in November 2022), specifically the share of writing and coding tasks among all required course tasks.[1] I use a difference-in-differences design to compare grade distributions before and after ChatGPT's release across courses with varying levels of AI task exposure. I find that grades rose substantially in high-exposure courses after 2022: the share of A grades increased by 13 percentage points (about 30% relative to the 2022 baseline) and GPA by 0.12 points, accompanied by compression of the grade distribution.

Grade increases in AI-exposed courses could reflect several mechanisms: genuine learning gains, sorting of stronger students into exposed courses, or grade inflation through task displacement. I distinguish between grade inflation and other mechanisms by differentiating between courses that rely more and less strongly on homework (versus in-class exams) in determining student grades. If grade increases reflect broad learning gains or ability sorting, they should appear regardless of whether assessment relies on homework or in-class exams. If instead they reflect task displacement, they should be concentrated in courses where homework carries greater assessment weight and AI output can substitute for student effort. Using a triple-differences (DDD) estimator that exploits variation in homework weight across courses with different AI exposure, I find that the AI exposure effect is significantly larger in courses with higher homework weight, which is difficult to reconcile with broad learning gains or sorting alone and is more consistent with task displacement as the primary mechanism.

The difference-in-differences results are robust to alternative definitions of AI task exposure, alternative enrollment thresholds for the balanced panel, and alternative pre-periods. A placebo test using the share of oral presentation tasks, where AI capabilities are weaker, finds no effect on grades, supporting the interpretation that the results are specific to tasks where AI capabilities are strongest. Taken together, these findings provide evidence that task displacement through generative AI is a novel mechanism of grade inflation, concentrated in courses whose task bundles overlap most strongly with AI capabilities and amplified when instructors cannot observe the production of graded work.

[1] Writing and coding are the domains where generative AI capabilities are strongest, making this measure a natural index of how strongly a course's assessed work overlaps with AI capabilities. Chirikov (2026) validates this measure, showing that it strongly predicts whether instructors adopted AI policies in their syllabi by 2025, confirming that it captures meaningful variation in AI pressure across courses and disciplines.

This paper makes three contributions. First, it contributes to the economics literature on grade inflation by providing causal evidence on a new mechanism distorting the informational value of grades: a technological shock to the production of graded work. Unlike mechanisms operating through grading rules, standards, or incentives, generative AI alters student output itself, at scale and with imperfect detectability. Second, it contributes to emerging research on generative AI and student achievement by showing that AI effects on grades are task specific. Grade increases are concentrated in courses whose task bundles overlap with AI capabilities, suggesting that measured gains may reflect changes in submitted work rather than underlying learning. Third, it extends task-based models of technological change from production to skill formation. While AI may affect learning through displacement, augmentation, or reinstatement of student task practice, this paper focuses on the grade-distortion implications of task displacement.

The findings have implications for the development and allocation of human capital. If grades become noisier signals of underlying skill in AI-exposed courses, employers and other users of transcripts may make less efficient matching decisions, increasing the need for alternative screening procedures and assessments of core skills. Distorted grade signals may also weaken human-capital development if students overestimate their mastery and underinvest in foundational skills. More broadly, if AI displaces skill-building tasks during learning, students may graduate with weaker capabilities in precisely the domains where AI is strongest, reinforcing a feedback loop between AI in education and AI in production that could accelerate automation. Together, these risks increase the importance of assessment reforms in education that preserve meaningful evaluation of skills in AI-exposed settings.

The paper proceeds as follows. Section 2 develops the conceptual framework and reviews the relevant literature. Section 3 describes the data, measures and methods. Section 4 presents the main results and robustness checks. Section 5 discusses implications for the informational value of grades and concludes.

2 Conceptual Framework

2.1 Grade Inflation: Consequences and Mechanisms

Grade inflation – rising grades unaccompanied by corresponding learning gains – is pervasive in higher education and has accelerated in recent decades. Denning et al. (2022) document rising GPAs across nine large public universities and national survey data, showing that increases cannot be explained by changes in student characteristics or preparation. The phenomenon is widespread: A has become the modal grade at many American universities, and the share of A grades has more than doubled at some institutions over the past two decades.[2] These trends matter because grades serve as signals of student skill used by students, employers, and graduate programs to make consequential decisions, and inflated grades have less informational value.

Grading standards shape student effort and achievement, with stricter standards raising performance and lenient grading reducing study time and weakening downstream preparation (Figlio and Lucas 2004; Babcock 2010; Gershenson et al. 2024). Grades also shape academic trajectories through belief formation and sorting (Elsner et al. 2021; McEwan et al. 2021). Grade inflation has measurable long-run consequences: school-level grade inflation affects earnings primarily through selection into higher-quality institutions and fields of study (Nordin et al. 2019), and teacher-level average-grade inflation lowers future test scores, college enrollment, and discounted lifetime earnings (Denning et al. 2026).

Prior research identifies several mechanisms through which grades become distorted signals. Instructor incentives tied to student evaluations and tenure generate strategic leniency, because higher grades improve evaluations and create incentives to inflate even absent learning gains (Langbein 2008; Carrell and West 2010; Keng 2018). Competitive signaling incentives across schools can make inflation contagious, as easy grades become strategic complements when schools compete for students and placements (Chan et al. 2007). Institutional grading and disclosure policies shape grade distributions in ways that can reduce comparability: mandatory grade caps alter course selection and field composition, while transcript reforms intended to provide more context can paradoxically push grades upward (Bar et al. 2012; Butcher et al. 2014). Finally, measurement error, discretion, and bias introduce noise into the grade signal independent of student performance, as teachers award higher grades through subjective judgment or

[2] At Harvard, for instance, the share of A grades rose from 24% in 2005 to 60.2% in 2025 (Claybaugh 2025).

systematically disadvantage students from certain groups (Hanna and Linden 2012; Diamond and Persson 2016). Despite their differences, all of these mechanisms share a common structure: they operate through how submitted work is evaluated, recorded, or disclosed, downstream of its production.

2.2 AI as a New Mechanism of Grade Inflation

The arrival of generative AI introduces a potentially new mechanism of grade inflation, distinct in kind from those established in the literature. AI systems now perform well on college assessments across a range of graded tasks (Geerling et al. 2023; Borges et al. 2024), and students increasingly use them for academic work, including to produce graded assignments (Contractor and Reyes 2025). Even with unchanged grading standards and unchanged instructor behavior, grades can rise when AI output substitutes for student effort on assessed tasks without any corresponding gain in student skill.

Empirical evidence on the impact of AI on grades is emerging but remains limited. The most relevant work is Hausman et al. (2025), who find at an Israeli university that AI availability raises grades and compresses grade distributions in courses with home-based assessments. The study compares within-student performance across courses with take-home and in-class assessments before and after ChatGPT's release, and provides suggestive evidence that grade increases reflect substitution rather than learning gains. These findings are informative and broadly consistent with the mechanism proposed here, but limited in two important ways. First, their exposure measure contrasts student performance in courses relying almost entirely on take-home assignments with courses relying almost entirely on in-class exams, and differences in pedagogy and assessment style between such distinct course types may confound the AI effect.[3] Second, their setting may not generalize beyond that institutional context, as the bimodal distribution of assessment formats they exploit is uncommon in many other higher education systems, including the United States, where courses typically combine both home and in-class assessment components. This paper develops a different approach, measuring AI exposure through course task composition, which is described in the next section.

[3] In particular, the post-COVID return to in-person instruction coincided with ChatGPT's release, and these course types may have adapted differently to that transition.

2.3 A Task-Based Framework

Task-based models of technological change conceptualize production as requiring the completion of a range of tasks, allocated between workers and technology based on comparative advantage (Autor et al. 2003; Acemoglu and Autor 2011). In this framework, technological change affects labor demand not by uniformly raising productivity but by reshaping the allocation of tasks between humans and technology. Automation displaces workers from tasks where technology has comparative advantage, reducing the labor share of those tasks, while the introduction of new tasks reinstates labor into activities where human comparative advantage is preserved (Acemoglu and Restrepo 2019).

This paper applies this logic to skill formation. In production, workers apply existing skills to perform tasks that generate output. In education, students develop skills by practicing cognitively demanding tasks. Different courses require different bundles of learning tasks. Research on learning shows that skill acquisition requires active engagement with challenging tasks rather than passive consumption of outputs, and that reducing the cognitive demands of practice can impair long-run skill development (Bjork and Bjork 2011; Soderstrom and Bjork 2015). AI affects task practice as a technology that can reshape which tasks students practice and how, with consequences not only for submitted work but for the skills students develop.

Applying task-based logic to education requires one conceptual extension. In production, the distinction between technology displacing workers and technology reinstating labor into new tasks is sufficient because what matters is output, not who produces it. In education, who performs the task determines what skills develop, which requires an additional distinction between AI replacing student effort and AI augmenting it without replacing it (Chirikov 2026).

This distinction motivates three mechanisms through which AI reshapes student task practice. *Task displacement* occurs when AI performs tasks instead of students, eliminating the practice opportunity through which skills develop. *Task augmentation* occurs when AI supports student performance of a task without replacing the essential cognitive effort, potentially enhancing learning by reducing peripheral burdens while preserving the core challenge. *Task reinstatement* occurs when AI enables entirely new tasks that were not previously feasible, creating new forms of practice and new skills.

While these mechanisms differ in their implications for skill development, they share a common prediction for grade distributions. Whether AI displaces student work by producing better

output than students would have achieved on their own, or augments and reinstates student task practice in ways that improve underlying skills, grades should rise. Moreover, grade increases should be larger in courses whose task bundles overlap more strongly with AI capabilities. Courses requiring more writing and coding are more exposed to AI than courses relying on oral presentations or laboratory work, and so should see larger grade increases after the arrival of generative AI.

Grade increases in AI-exposed courses are therefore consistent with multiple mechanisms and do not by themselves identify whether displacement or other channels are responsible. The distinction matters because only displacement erodes the informational value of grades: if AI improves submitted work without improving underlying skills, grades become less reliable signals of student capability. Distinguishing among mechanisms therefore requires additional variation. Under augmentation or reinstatement, AI improves underlying student skill, which should manifest across all forms of assessment, homework and in-class exams alike. Sorting of stronger students into more AI-exposed courses after ChatGPT's release would similarly produce grade increases across assessment types. Under displacement, by contrast, AI output substitutes directly for unsupervised student work, so grade increases should be concentrated in courses where homework carries greater assessment weight. In-class assessments, where students must perform without AI assistance, are insulated from displacement even in highly AI-exposed courses. The distribution of grade increases across assessment formats therefore helps distinguish displacement from other channels.

3 Data, Measures and Methods

3.1 Data

3.1.1 Setting

The study uses data from a large, selective public research university in Texas enrolling over 50,000 students across all major academic fields. Two features of this setting are essential for the analysis. First, Texas state law requires all public universities to post undergraduate course syllabi online, making course-level task composition and grade components observable. Second, the university issued no institutional mandate related to the use of AI tools by students, so the course-level responses documented by my earlier analysis (Chirikov 2026) and the grade changes

analyzed here reflect organic faculty and student behavior rather than centrally administered policy.

3.1.2 Grade Data

The university publicly posts semester-level grade distributions at the course-section level. For each section in each term, the data reports the number of students receiving each letter grade from A through F. I use Fall and Spring semesters from 2018 through 2025, excluding Summer sessions. The grade data are subject to confidentiality suppression rules: a course-section does not appear in the published distributions if it has fewer than 10 undergraduate students, if all or all-but-one students received the same grade, or if no students received a standard letter grade (e.g., the course was graded entirely on a pass/fail basis). These suppression rules have little practical consequence for the analysis, as grade distributions in very small or non-standard courses are too noisy to support meaningful inference about grade inflation.

3.1.3 Syllabi

Course syllabi for the same institution were collected by web-scraping the university's public course catalog from 2018 through 2025, following the procedure described in Chirikov (2026). For each course listing, the scrape extracted year, semester, department, course number, course title, instructor names, and the full syllabus PDF. The university departments are classified into nine broad disciplinary fields: Arts, Business, Math and Computer Science (CS), Engineering, Humanities, Life Sciences, Physical Sciences, Social Sciences, and Other. In this paper, syllabi serve two purposes: they are the source of AI task exposure measures and course assessment weights, specifically the percentage weight assigned to take-home assignments. The construction of both measures is described in detail in Section 3.2.

3.1.4 Analytical Sample

From the linked grade and syllabus data I construct a balanced panel of courses observed in every Fall semester from 2018 through 2025, where a course is identified by department, course number, and title. The panel is restricted to Fall semesters for two reasons. First, restricting to a single semester per year removes within-year seasonality in grade distributions. Second, the COVID-19 pandemic disrupted only one Fall semester (Fall 2020) compared to two Spring semesters (Spring 2020 and Spring 2021), so the Fall-only panel provides a cleaner pre-treatment baseline. I retain Fall 2020 in the main sample and show robustness to dropping it in Table A6.

A course is included if it appears in every Fall from 2018 through 2025 with more than 20 enrolled students in each year.[4] The resulting panel contains 319 courses spanning 84 departments and 2,552 course-year observations over eight Fall semesters, covering 507,076 student-course enrollments in total. The sample covers all nine disciplinary fields, with Business (21%), Social Sciences (20%), Engineering (15%), Humanities (14%), and Math/CS (13%) making up the largest shares. Mean enrollment per course-year is 199 students and the median is 98, reflecting the mix of large lecture courses and smaller upper-division offerings in the panel. Table 1, Panel A provides the full sample description.

3.2 Measures

3.2.1 Grade Outcomes

The main outcomes are constructed from the course-level grade distributions described in Section 3.1.2. For each course-year observation, I compute the share of students receiving each letter grade as a fraction of total enrolled students. The university does not award A+ grades, so the highest grade is A. Grade shares are computed for A, A-, B+, B, B-, C, and D/F categories. Given the small number of students in each subcategory, C+, C, and C- grades are combined into a single share of C grades, and D+, D, D-, and F grades are combined into a single share of D and F grades.

The primary summary outcome is the share of A grades. I also compute two aggregate measures: mean GPA, calculated according to the university's standard grade point scale (A=4.0, A-=3.67, and so on), and the within-course standard deviation of GPA, which captures compression of the grade distribution. A reduction in GPA standard deviation alongside an increase in mean GPA indicates that grades are concentrating at the top of the distribution rather than shifting uniformly upward.

In addition to individual grade shares, I construct a set of cumulative threshold outcomes: the share of students receiving a given grade or better (at least A-, at least B+, at least B, at least B-, and at least C-). These cumulative outcomes characterize how the entire grade distribution shifts at different points on the scale, complementing the individual grade share estimates.

Table 1, Panel B reports descriptive statistics for all grade outcomes in Fall 2022 and Fall 2025. In Fall 2022, the mean share of A grades across courses was 0.44 and mean GPA was 3.40,

[4] The threshold of 20 students ensures that grade distributions are sufficiently stable to support meaningful inference: with very small enrollments, a single student's grade can substantially shift the distribution, introducing measurement noise into the outcomes. Results are robust to varying this threshold between 15 and 40 students (Table A5).

with a within-course GPA standard deviation of 0.74. By Fall 2025, the mean share of A grades had risen to 0.48, mean GPA to 3.49, and GPA standard deviation had fallen to 0.66, indicating both upward drift and compression of the grade distribution over the panel period. Shares of B and lower grades declined correspondingly, with the share of D and F grades falling from 0.033 to 0.024.

3.2.2 AI Task Exposure

The main exposure measure captures the share of a course's required tasks that fall in domains where generative AI capabilities are strongest, such as writing and coding. Tasks are measured from the Fall 2022 syllabus of each course in the panel to ensure that the measure is determined prior to the treatment.[5]

Task extraction uses a large language model (GPT-4.1-mini) that reads each syllabus and identifies all concrete academic activities required of students using a verb-object structure (e.g., "write essay," "code simulation," "read paper," "present findings"). The extraction prompt explicitly excludes administrative actions and abstract course goals. To reduce noise from verbosity, semantically equivalent task mentions within a syllabus are deduplicated, so that "write a 2-page essay" and "write a 5-page essay" count as a single writing task rather than two.

The main exposure variable is the share of writing and coding tasks among all deduplicated tasks, which ranges from 0 to 1 with a mean of 0.24 and standard deviation of 0.21 (Table 1, Panel C). Robustness checks use alternative operationalizations: the share computed from all task mentions rather than deduplicated tasks, separate writing-only and coding-only shares, and a binary indicator for any writing or coding task. I also extract the share of oral presentation tasks as a placebo measure: AI capabilities for oral presentations are substantially weaker, so this measure should not predict grade changes if the main results are specific to AI-capable tasks.

The construction and validity of the AI exposure measure are documented in Chirikov (2026). Task composition varies across disciplines in expected ways, with writing tasks most prevalent in Humanities and Social Sciences, coding tasks concentrated in Engineering and Math/CS, and design tasks highest in Arts. Courses with higher shares of writing and coding tasks are also

[5] Fall 2022 syllabi were published in August 2022, before ChatGPT's release on November 30. ChatGPT became available only twelve days before Fall 2022 exams ended, so any limited use during that period would likely have moved Fall 2022 grades in the same direction as later AI use, making the estimates more conservative.

substantially more likely to adopt AI policies by Fall 2025, consistent with instructors in high-exposure courses recognizing the relevance of AI to their assessments.

3.2.3 Assessment Weights

The assessment weight measure captures the share of a course's final grade allocated to take-home assignments (including unsupervised take-home exams and quizzes), as opposed to supervised exams and in-class work. It is extracted from the same Fall 2022 syllabi used for task exposure measurement using a similar LLM-assisted classification procedure described above. The main variable is the share of the course grade allocated to homework and take-home assignments, which ranges from 0 to 100% with a mean of 37% and a median of 30% (Table 1, Panel C). For the eight courses (less than 3%) where assessment weights were not present in the syllabus, I impute the discipline-level mean from observed courses in the same Fall 2022 cross-section. A robustness check restricts the sample to courses with observed weights only, leaving results unchanged (Table A13).

Table 1. Descriptive Statistics

Panel A. Sample structure

Statistic	Value
Unique courses	319
Course-year observations	2,552
Student-course enrollments	507,076
Unique departments	84
Years in panel	2018-2025
Minimum course size rule	> 20 every Fall year
Mean students per course-year	198.7
Median students per course-year	98
Courses with AI policy in 2025 (%)	62.7
Discipline: Arts (%)	3.76
Discipline: Business (%)	20.69
Discipline: Math/CS (%)	12.85
Discipline: Engineering (%)	15.05
Discipline: Humanities (%)	14.42
Discipline: Life Sciences (%)	8.15
Discipline: Other (%)	1.25
Discipline: Physical Sciences (%)	4.08
Discipline: Social Sciences (%)	19.75

Panel B. Outcome variables

Outcome-Year	Mean	SD	P25	Median	P75	Min	Max	N
2022 Share of A	0.4398	0.2005	0.2857	0.3977	0.5832	0.0149	0.9583	319
2025 Share of A	0.4775	0.2186	0.3085	0.4507	0.6612	0.0303	0.9317	319
2022 Share of A-	0.1453	0.0869	0.0847	0.1348	0.1922	0.0000	0.5207	319
2025 Share of A-	0.1429	0.0847	0.0878	0.1282	0.1933	0.0000	0.4286	319
2022 Share of B+	0.0970	0.0578	0.0563	0.0958	0.1319	0.0000	0.3409	319
2025 Share of B+	0.0941	0.0657	0.0484	0.0809	0.1277	0.0000	0.4110	319
2022 Share of B	0.1024	0.0766	0.0537	0.0925	0.1343	0.0000	0.5000	319
2025 Share of B	0.0919	0.0706	0.0382	0.0788	0.1231	0.0000	0.3654	319
2022 Share of B-	0.0508	0.0390	0.0229	0.0448	0.0714	0.0000	0.2000	319
2025 Share of B-	0.0444	0.0399	0.0112	0.0378	0.0687	0.0000	0.2553	319
2022 Share of C	0.0841	0.0732	0.0269	0.0617	0.1204	0.0000	0.3364	319
2025 Share of C	0.0685	0.0727	0.0143	0.0417	0.1017	0.0000	0.4848	319
2022 Share of D and F	0.0329	0.0343	0.0000	0.0241	0.0494	0.0000	0.2083	319
2025 Share of D and F	0.0240	0.0303	0.0000	0.0145	0.0357	0.0000	0.2059	319
2022 Share of Other	0.0476	0.0420	0.0156	0.0382	0.0714	0.0000	0.2574	319
2025 Share of Other	0.0566	0.0629	0.0130	0.0370	0.0774	0.0000	0.4760	319
2022 GPA	3.4046	0.3072	3.2059	3.4408	3.6256	2.4468	3.9631	319
2025 GPA	3.4876	0.3111	3.2942	3.5340	3.7420	2.3969	3.9620	319
2022 GPA SD	0.7436	0.2337	0.5655	0.7337	0.9288	0.1316	1.2517	319
2025 GPA SD	0.6556	0.2390	0.4610	0.6414	0.8255	0.1068	1.3148	319

Panel C. Exposure and homework weight variables

Variable	Mean	SD	P25	Median	P75	Min	Max	N
Share of writing + coding tasks	0.239	0.208	0.059	0.2	0.364	0	1.000	319
Share of writing tasks	0.227	0.207	0.000	0.2	0.353	0	1.000	319
Share of coding tasks	0.012	0.051	0.000	0.0	0.000	0	0.385	319
Any writing task	0.730	0.444	0.000	1.0	1.000	0	1.000	319
Any coding task	0.066	0.248	0.000	0.0	0.000	0	1.000	319
Has AI policy in 2025	0.627	0.484	0.000	1.0	1.000	0	1.000	319
Homework weight	0.367	0.251	0.200	0.3	0.550	0	1.000	319

3.3 Methods

3.3.1 Grade Effects and Mechanism

To estimate the effect of AI exposure on grade distributions, I use a difference-in-differences design that compares grade changes before and after ChatGPT's release across courses with varying levels of AI task exposure. The main estimating equation is:

$$Y_{ct} = \alpha_c + \lambda_t + \beta\,(\text{Post}_t * \text{Exposure}_c) + \varepsilon_{ct} \quad (1)$$

where Y_{ct} is a grade outcome for course c in year t, α_c is a course fixed effect, λ_t is a year fixed effect, $\text{Post}_t = 1$ for $t >= 2023$, and Exposure_c is the share of writing and coding tasks measured from the Fall 2022 syllabus. The coefficient β captures the differential post-ChatGPT change in grade outcomes for courses with higher AI task exposure. Standard errors are clustered at the course level throughout; department-level clustering is reported as a robustness check (Table A2). I also estimate a dynamic version of equation (1) by replacing Post_t * Exposure_c with a full set of year-by-exposure interactions, normalizing to 2022. This produces the event-study figures and is used for the pre-trend tests.

The identifying assumption is that, absent ChatGPT's release, grade trends would have evolved in parallel across high- and low-exposure courses. Several features of the design support this assumption. The exposure measure is constructed from pre-treatment syllabi, removing any possibility of reverse causation. The balanced panel holds course composition fixed across years. I test the assumption using a joint Wald test of the pre-treatment year-by-exposure interactions from the dynamic specification (Table A3).

Grade increases in AI-exposed courses are consistent with multiple mechanisms: genuine learning gains through augmentation or reinstatement, sorting of stronger students into exposed courses, or grade inflation through task displacement. As developed in Section 2, these mechanisms have different predictions for where grade increases should be concentrated. Under augmentation/reinstatement or sorting, grades should rise regardless of assessment format. Under displacement, AI can substitute for student effort on take-home work, but is less likely to assist with supervised in-class assessments, so grade increases should be larger in courses where homework carries greater weight.

To distinguish between these mechanisms, I augment equation (1) with a triple interaction that exploits variation in homework weight:

$$Y_{ct} = \alpha_c + \lambda_t + \beta_1\left(\text{Post}_t * \text{Exposure}_c\right) + \beta_2\left(\text{Post}_t * \text{HighHW}_c\right) + \beta_3(\text{Post}_t * \text{Exposure}_c * \text{HighHW}_c) + \varepsilon_{ct} \quad (2)$$

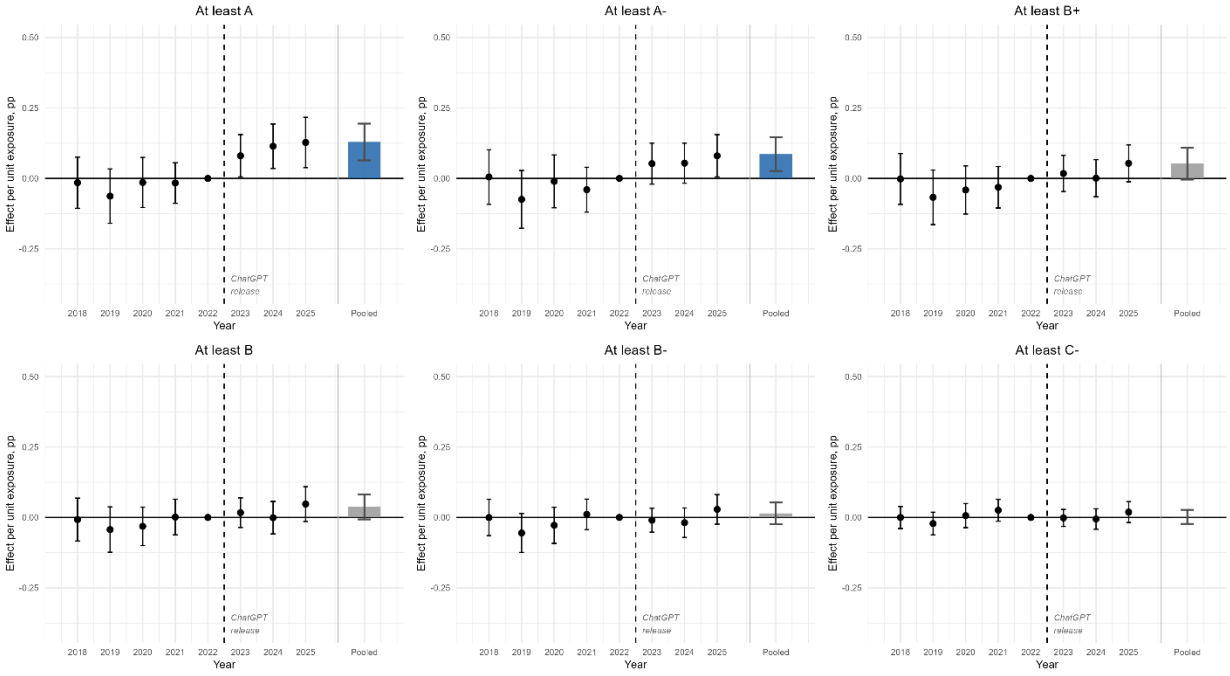
where HighHW_c is an indicator for above-median homework weight (median = 30%), defined from the Fall 2022 cross-section and fixed across years. The coefficient β_1 captures the post-ChatGPT exposure effect in low-homework courses, and β_3 captures the additional effect in high-homework courses. A positive and significant β_3 is consistent with task displacement and difficult to reconcile with augmentation or sorting alone, both of which predict $\beta_3 \approx 0$. I also estimate a dynamic version of equation (2) to examine pre-trends in the triple interaction (Table A12) and produce the event-study figures for the mechanism test (Figure 2).

4 Results

4.1 The Impact of AI Exposure on Grades

Figure 1 and corresponding Table A1 present event-study and pooled estimates for cumulative grade thresholds. Pre-treatment coefficients are close to zero and jointly insignificant across all outcomes (Table A3), supporting the parallel trends assumption. After ChatGPT's release, the share of students receiving an A increased substantially in high-exposure courses, with effects diminishing and becoming insignificant further down the distribution. The pooled estimate for the share of students receiving an A is 13 percentage points, or approximately 30% relative to the 2022 baseline mean of 0.44. The effect becomes smaller and less significant further down the distribution: 9 percentage points for at least A-, 5 percentage points for at least B+, and small and insignificant below that. This pattern is consistent with AI primarily converting A- and B+ grades into A grades, with little effect further down the distribution.

Figure 1. Grade Effects by AI Exposure (Cumulative Thresholds)



Notes: Each panel shows event-study estimates from equation (1) for cumulative grade thresholds, defined as the share of students receiving a given grade or better. The dashed vertical line marks ChatGPT's release in November 2022. The bar to the right of each panel shows the pooled estimate from the same specification. Bars are colored blue if the pooled estimate is positive and significant at the 5% level, red if negative and significant, and gray otherwise. Whiskers show 95% confidence intervals. Standard errors clustered at the course level.

Table 2 and corresponding Figure A1 present event-study and pooled estimates for individual grade shares and summary measures. The share of A- grades fell by 4 percentage points, the share of B+ grades by 3 percentage points, and the shares of B and below show smaller and mostly insignificant changes. Mean GPA rose by 0.12 points, and the within-course GPA standard deviation fell by 0.09 points, indicating that grades are not shifting uniformly upward but concentrating at the top of the distribution. The shares of D and F grades, and shares of other grades (withdrawals, incompletes, and credit/no credit) do not change. Course enrollments are also unaffected, providing no evidence of differential changes in course size or aggregate demand for AI-exposed courses after ChatGPT's release.

Table 2. Grade Effects by AI Exposure

Outcome	Effect (SE)
Share of A	0.13 (0.03)***
Share of A-	-0.04 (0.01)***
Share of B+	-0.03 (0.01)***
Share of B	-0.01 (0.01)
Share of B-	-0.02 (0.01)***
Share of C	-0.01 (0.01)
Share of D and F	-0.01 (0.01)
Share of Other	0.00 (0.01)
GPA	0.12 (0.05)**
GPA SD	-0.09 (0.04)**
Enrollment	19.85 (27.92)
Observations	2,552
Courses	319

Notes: Pooled post-treatment estimates from equation (1). Standard errors clustered at the course level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

4.2 Mechanism: Task Displacement vs Learning Gains and Sorting

The observed grade increases are consistent with multiple mechanisms. To distinguish between task displacement and other mechanisms – learning gains through augmentation or reinstatement, and sorting of stronger students into high-exposure courses – I exploit variation in homework weight across courses using equation (2).[6] Under displacement, grade increases should be concentrated in courses where homework carries greater assessment weight, since AI can substitute for student task practice on unsupervised take-home work but is less likely to assist with supervised in-class assessments. Under augmentation or sorting, grade increases should appear regardless of assessment format, predicting a near-zero triple interaction coefficient β_3.

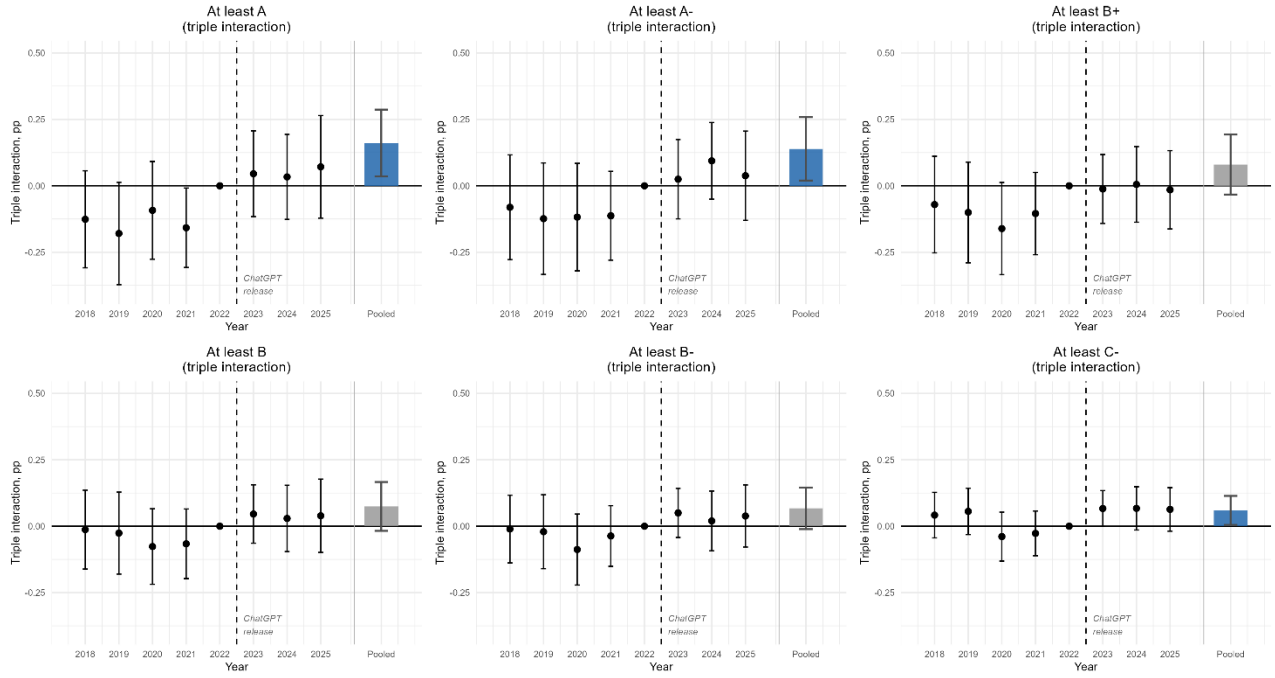
Figure 2 and Table 3 (and corresponding Table A10 and Figure A2) present the DDD estimates. For courses with below-median homework weight, the post-treatment exposure effect

[6] One potential concern is that AI task exposure and homework weight may be collinear: courses requiring more writing and coding may also tend to rely more on take-home assignments, which could make it difficult to separately identify the exposure effect and the homework moderator. Table A9 addresses this by adding post * homework weight as a direct control to equation (1), using both a continuous homework weight and a binary above-median indicator. The main exposure coefficient is stable across all three specifications, confirming that the exposure and homework weight dimensions are sufficiently distinct to support the DDD design.

is small and insignificant for the share of A grades and GPA. The triple interaction β_3 is positive and significant for the share of A grades: above-median homework courses show an additional 16 percentage point increase in the share of A grades relative to below-median courses with the same level of AI exposure. The effect on GPA follows a similar pattern, with an additional 0.13 point increase in high-homework courses, though this estimate is less precisely estimated. The triple interaction is insignificant for enrollment, suggesting no differential changes in the size of high-homework, high-exposure courses after ChatGPT's release.

The share of other grades shows an additional pattern worth noting. In above-median homework courses, the share of other grades declines significantly with AI exposure relative to below-median homework courses. This suggests that AI in homework-heavy courses may convert what would otherwise have been non-standard grade outcomes into letter grades, further contributing to the grade increases documented above.

Figure 2. Grade Effects by AI Exposure and Homework Weight (Cumulative Thresholds)



Notes: Each panel shows event-study estimates of the triple interaction term from equation (2) for cumulative grade thresholds, defined as the share of students receiving a given grade or better. The dashed vertical line marks ChatGPT's release in November 2022. The bar to the right of each panel shows the pooled triple interaction estimate from the same specification. Bars are colored blue if the pooled estimate is positive and significant at the 5% level, red if negative and significant, and gray otherwise. Whiskers show 95% confidence intervals. Standard errors clustered at the course level.

Table 3. Grade Effects by AI Exposure and Homework Weight

Outcome	Post x Exposure (low HW)	Post x Exposure x High HW
Share of A	0.017 (0.046)	0.161 (0.064)**
Share of A-	-0.022 (0.022)	-0.022 (0.030)
Share of B+	0.009 (0.016)	-0.059 (0.024)**
Share of B	-0.012 (0.021)	-0.006 (0.027)
Share of B-	-0.017 (0.013)	-0.007 (0.016)
Share of C	-0.009 (0.018)	-0.008 (0.023)
Share of D and F	-0.004 (0.010)	-0.004 (0.013)
Share of Other	0.039 (0.019)**	-0.056 (0.022)***
GPA	0.044 (0.073)	0.132 (0.099)
GPA SD	-0.106 (0.047)**	0.031 (0.070)
Enrollment	-24.481 (38.600)	58.605 (51.552)
Observations	2,552	2,552
Courses	319	319

Notes: Pooled post-treatment estimates from equation (2). Standard errors clustered at the course level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

The concentration of grade increases in high-homework, high-exposure courses is difficult to reconcile with augmentation or sorting. If AI were improving underlying student skills, those gains should manifest in both homework and exam performance, producing grade increases regardless of assessment format. If stronger students were sorting into high-exposure courses after ChatGPT's release, that sorting should similarly affect grades across all assessment types. The pattern instead points to task displacement: AI substituting for student effort specifically on the unsupervised assessments where instructors cannot observe the production of submitted work.

4.3 Robustness

The main results are robust to alternative exposure definitions, panel construction choices, standard error clustering, and additional controls.

The positive effect on the share of A grades and the negative effects on lower grade shares are stable across alternative definitions of AI task exposure (Table A4): all task mentions rather than deduplicated tasks, a binary indicator for any writing or coding task, and writing-only share. Coding-only estimates are directionally similar but less precisely estimated, reflecting the smaller number of coding-intensive courses in the sample. The placebo test using the share of oral

presentation tasks yields no significant effects on any grade outcome (Table A7), supporting the interpretation that the results are specific to tasks where AI capabilities are strongest.

The results are also insensitive to the panel construction choices. Table A5 shows that the main estimates are stable when varying the minimum enrollment threshold between 15 and 40 students. Table A6 shows that results are unchanged when dropping Fall 2020 and when restricting to a shorter 2021-2025 pre-period, indicating that the COVID-affected semester and the length of the pre-period do not drive the findings.

The main estimates remain significant when clustering standard errors at the department rather than the course level (Table A2), and are unchanged when adding instructor stability (the share of Fall 2022 instructors teaching the course in each year) as a time-varying control (Table A8), suggesting that instructor turnover is not confounding the results.

The DDD results are similarly robust. The triple interaction coefficient itself is stable when removing 8 courses with imputed homework weights, when using the mean rather than the median as the high-homework threshold, and across stricter enrollment thresholds of 30 and 40 students (Table A13). The triple interaction also remains significant when clustering standard errors at the department level (Table A11).

5 Discussion

This paper provides novel evidence that generative AI has become a new mechanism of grade inflation in higher education. I find that courses with higher shares of writing and coding tasks saw substantial grade increases after ChatGPT's release, accompanied by compression of the grade distribution. These effects are concentrated in courses where homework carries greater assessment weight, consistent with AI substituting for student effort on unsupervised take-home work rather than improving underlying skills.

This paper makes three contributions. First, it contributes to the literature on grade inflation by identifying a new mechanism through which grades become distorted signals of student skill. Prior research has documented grade inflation across institutions and identified mechanisms operating through instructor incentives (Langbein 2008; Carrell and West 2010; Keng 2018), signaling competition (Chan et al. 2007), institutional policies (Bar et al. 2012; Butcher et al. 2014), and measurement error and bias (Hanna and Linden 2012; Diamond and Persson 2016).

These mechanisms share a common structure: they operate downstream of the production of student work, through how submitted work is evaluated, recorded, or disclosed. Generative AI introduces a fundamentally different channel: it alters the production of graded work itself, before instructors observe and assess it. The resulting distortion is selective rather than uniform, concentrated in courses whose task bundles overlap most strongly with AI capabilities and amplified where assessment relies on unsupervised work. This selectivity has implications beyond the average grade level: it reduces the comparability of grades across courses and potentially across students with unequal access to more capable AI tools, eroding the informational value of transcripts in ways that are difficult to detect from grade distributions alone.

Second, the paper contributes to the emerging literature on the impact of generative AI on student achievement. Existing work shows that AI systems perform well on college assessments (Geerling et al. 2023; Borges et al. 2024) and that students increasingly use them for academic work, including to produce graded assignments (Contractor and Reyes 2025). This paper provides evidence that the effects of AI on grades are task-specific, concentrated in courses whose task bundles overlap with AI capabilities, and amplified by assessment formats that give AI the greatest scope to substitute for student effort. It also cautions that measured gains in student performance following AI adoption may reflect shifts in the production of submitted work rather than genuine improvements in underlying skill – a distinction that has direct implications for how AI-era grade data should be interpreted by employers, graduate programs, and researchers studying human capital formation. These findings extend recent evidence from Hausman et al. (2025), who find that AI availability raises grades in courses with home-based assessments, by showing that the effect is driven by task composition rather than assessment format alone.

Third, the paper also demonstrates the potential of a task-based approach to study the impact of AI on skill formation in educational settings. While prior work focuses on the impact of AI on production, occupations, and worker productivity (Webb 2019; Eloundou et al. 2024; Brynjolfsson et al. 2025), this paper shows that task-based models also provide a useful framework for understanding how AI operates in educational settings. Applied to higher education, the framework highlights that students develop skills by completing learning tasks, and that AI can affect learning by displacing, augmenting, or reinstating student task practice (Chirikov 2026). The grade-distortion implications documented here follow directly from the displacement mechanism: when

AI improves submitted work on assessed tasks without corresponding skill growth, grades become less informative measures of student capability.

Several limitations of the study are worth discussing. First, the analysis covers a single institution, which limits external generalizability. However, the findings are consistent with a general mechanism that is not specific to this institution, and the breadth of disciplinary coverage in the sample suggests the results are unlikely to reflect idiosyncratic features of a single field or department.

Second, the balanced panel is restricted to courses large enough to appear in the public grade distributions, which may underrepresent small seminars and upper-division courses. Varying the enrollment threshold between 15 and 40 students leaves results unchanged, suggesting this restriction does not drive the findings.

Third, the exposure measure captures task composition from syllabi but does not directly measure actual student AI use at the course level. However, prior research showed that courses with higher task exposure are substantially more likely to have adopted explicit AI policies (Chirikov 2026), indicating that instructors in high-exposure courses independently recognized the relevance of AI to their assessments.

Fourth, the findings describe average effects across courses and students, but there is likely substantial heterogeneity in how individual students engage with AI that the design cannot capture. Some students may be using AI to substitute for their own effort on graded work, while others may be using it to augment their learning or to practice new AI-related tasks through reinstatement. The aggregate grade increases documented here reflect the net effect of these individual-level differences, and the DDD identifies displacement as the dominant mechanism on average, but the distribution of mechanisms across students within a course remains unobserved. Understanding this heterogeneity requires individual-level data linking student AI use to performance across different assessment formats, which remains an important direction for future research.

The findings have implications for the development and allocation of human capital. If grades in writing- and coding-intensive courses increasingly reflect AI-assisted output rather than student skill, employers and graduate programs relying on transcripts for screening and matching decisions may make less efficient allocations, increasing the demand for alternative assessments of core capabilities. Distorted grade signals may also weaken human capital development if students, receiving higher grades, overestimate their mastery and underinvest in foundational skills. More

broadly, if AI displaces skill-building tasks during learning, students may graduate with weaker capabilities in precisely the domains where AI is strongest, reinforcing a feedback loop between AI in education and AI in production that could accelerate automation and widen skill gaps in the labor market.

Assessment reform is the most direct institutional response, but its design is non-trivial (Corbin et al. 2025). Moving all assessments to supervised in-person environments would limit AI's scope to substitute for student work, but this approach has important constraints. Not all skills can be meaningfully evaluated under exam conditions: the ability to produce a well-researched essay, develop a software project, or conduct an empirical analysis requires sustained engagement that timed in-person assessments cannot capture. Restricting assessment to formats that are AI-proof risks measuring a narrower set of capabilities than the ones courses are designed to develop, potentially undermining the learning goals that graded work is meant to serve. A more promising direction is to redesign assessments so that AI use is either structurally constrained by the task or purposefully incorporated into it, for example, by requiring students to document their process, justify their choices, or demonstrate understanding through follow-up interaction. The evidence in this paper that instructors in high-exposure courses are already moving toward more explicit AI regulation (Chirikov 2026) suggests that course-level adaptation is underway, but the pace and effectiveness of these responses remain open questions for future research.

References

1. Acemoglu, D., & Autor, D. H. (2011). Skills, tasks and technologies: Implications for employment and earnings. In O. Ashenfelter & D. Card (Eds.), Handbook of labor economics (Vol. 4, pp. 1043-1171). Elsevier.
2. Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. Journal of Economic Perspectives, 33(2), 3-30.
3. Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. Quarterly Journal of Economics, 118(4), 1279-1333.
4. Babcock, P. (2010). Real costs of nominal grade inflation? New evidence from student course evaluations. Economic Inquiry, 48(4), 983-996.
5. Bar, T., Kadiyali, V., & Zussman, A. (2012). Putting grades in context. Journal of Labor Economics, 30(2), 445-478.
6. Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher et al. (Eds.), Psychology and the real world (pp. 56-64). Worth Publishers.
7. Bleemer, Z., & Mehta, A. (2024). College major restrictions and student stratification (No. w33269). National Bureau of Economic Research.
8. Borges, B., Foroutan, N., Bayazit, D., Sotnikova, A., Montariol, S., Nazaretsky, T., ... & EPFL Data Consortium. (2024). Could ChatGPT get an engineering degree? Evaluating higher education vulnerability to AI assistants. Proceedings of the National Academy of Sciences, 121(49).
9. Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. Quarterly Journal of Economics, 140(2), 889-942.
10. Butcher, K. F., McEwan, P. J., & Weerapana, A. (2014). The effects of an anti-grade-inflation policy at Wellesley College. Journal of Economic Perspectives, 28(3), 189-204.
11. Carrell, S. E., & West, J. E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. Journal of Political Economy, 118(3), 409-432.
12. Chan, W., Li, H., & Suen, W. (2007). A signaling theory of grade inflation. International Economic Review, 48(3), 1065-1090.
13. Chirikov, I. (2026). How Instructors Regulate AI in College: Evidence from 31,000 Course Syllabi. Higher Education Working Paper Series, 26(1).

14. Claybaugh, A. (2025). Re-Centering academics at Harvard College: Update on grading and workload (October 2025). Office of Undergraduate Education, Harvard College.
15. Contractor, Z., & Reyes, G. (2025). Generative AI in higher education: Evidence from an elite college. IZA Discussion Paper, (18055).
16. Corbin, T., Dawson, P., Nicola-Richmond, K., & Partridge, H. (2025). 'Where's the line? It's an absurd line': towards a framework for acceptable uses of AI in assessment. Assessment & Evaluation in Higher Education, 50(5), 705-717.
17. Denning, J. T., Eide, E. R., Mumford, K. J., Patterson, R. W., & Warnick, M. (2022). Why have college completion rates increased?. American Economic Journal: Applied Economics, 14(3), 1-29.
18. Denning, J. T., Nesbit, R. L., Pope, N. G., & Warnick, M. (2026). Easy A's, less pay: The long-term effects of grade inflation. NBER Working Paper 34952.
19. Diamond, R., & Persson, P. (2016). The long-term consequences of teacher discretion in grading of high-stakes tests (No. w22207). National Bureau of Economic Research.
20. Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. Science, 384(6702), 1306-1308.
21. Elsner, B., Isphording, I. E., & Zölitz, U. (2021). Achievement rank affects performance and major choices in college. Economic Journal, 131(640), 3182-3206.
22. Figlio, D. N., & Lucas, M. E. (2004). Do high grading standards affect student performance? Journal of Public Economics, 88(9-10), 1815-1834.
23. Geerling, W., Mateer, G. D., Wooten, J., & Damodaran, N. (2023). ChatGPT has aced the test of understanding in college economics: Now what?. The American Economist, 68(2), 233-245.
24. Gershenson, S., Holt, S. B., & Tyner, A. (2024). Making the grade: The effect of teacher grading standards on student outcomes. Contemporary Economic Policy, 42(2), 305-318.
25. Hanna, R. N., & Linden, L. L. (2012). Discrimination in grading. American Economic Journal: Economic Policy, 4(4), 146-168.
26. Hausman, N., Rigbi, O., & Weisburd, S. (2025). Generative AI's Impact on Student Achievement and Implications for Worker Productivity. Available at SSRN 5393516.
27. Keng, S. H. (2018). Tenure system and its impact on grading leniency, teaching effectiveness and student effort. Empirical Economics, 55(3), 1207-1227.

28. Langbein, L. (2008). Management by results: Student evaluation of faculty teaching and the mis-measurement of performance. Economics of Education Review, 27(4), 417-428.
29. McEwan, P. J., Rogers, S., & Weerapana, A. (2021, May). Grade sensitivity and the economics major at a women's college. In AEA papers and proceedings (Vol. 111, pp. 102-106).
30. Nordin, M., Heckley, G., & Gerdtham, U. (2019). The impact of grade inflation on higher education enrolment and earnings. Economics of Education Review, 73, 101936.
31. Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. Perspectives on Psychological Science, 10(2), 176-199.
32. Webb, M. (2019). The impact of artificial intelligence on the labor market. Available at SSRN 3482150.

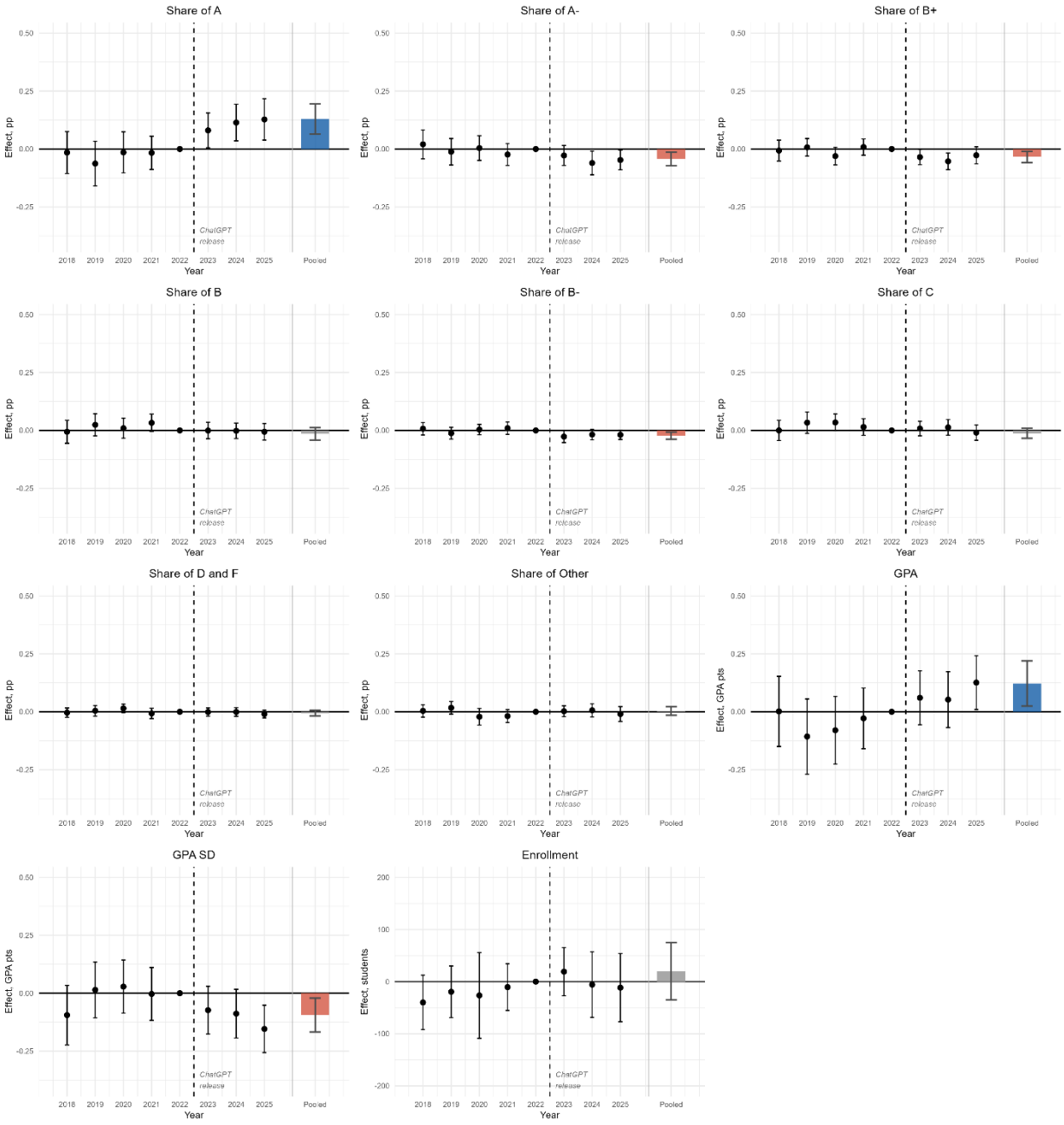
Appendix

Table A1. Grade Effects by AI Exposure (Cumulative Thresholds)

Outcome	Effect (SE)
At least A	0.13 (0.03)***
At least A-	0.09 (0.03)***
At least B+	0.05 (0.03)*
At least B	0.04 (0.02)*
At least B-	0.01 (0.02)
At least C-	0.00 (0.01)
Observations	2,552
Courses	319

Notes: Pooled post-treatment estimates from equation (1) that are shown in Figure 1. Standard errors clustered at the course level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Figure A1. Grade Effects by AI Exposure



Notes: Each panel shows event-study estimates from equation (1) for the share of a given grade and summary measures reported in Table 2. The dashed vertical line marks ChatGPT's release in November 2022. The bar to the right of each panel shows the pooled estimate from the same specification. Bars are colored blue if the pooled estimate is positive and significant at the 5% level, red if negative and significant, and gray otherwise. Whiskers show 95% confidence intervals. Standard errors clustered at the course level.

Table A2. Grade Effects by AI Exposure (Department-Clustered Standard Errors)

Outcome	Effect (SE)
Share of A	0.129 (0.034)***
Share of A-	-0.043 (0.012)***
Share of B+	-0.034 (0.013)**
Share of B	-0.015 (0.013)
Share of B-	-0.023 (0.008)***
Share of C	-0.013 (0.011)
Share of D and F	-0.005 (0.006)
Share of Other	0.004 (0.011)
GPA	0.122 (0.052)**
GPA SD	-0.094 (0.034)***
Enrollment	19.853 (29.198)
Observations	2,552
Courses	319

Notes: Pooled post-treatment estimates from equation (1). Standard errors clustered at the department level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Table A3. Pre-Trend Tests by AI Exposure

Outcome	Chi-square	p-value
Share of A	2.209	0.697
Share of A-	3.677	0.451
Share of B+	4.630	0.327
Share of B	5.792	0.215
Share of B-	3.324	0.505
Share of C	6.153	0.188
Share of D and F	8.101	0.088
Share of Other	9.152	0.057
GPA	4.699	0.320
GPA SD	5.970	0.201
Enrollment	3.150	0.533
Observations	2,552	
Courses	319	

Notes: Joint Wald tests of whether the 2018-2021 year-by-exposure coefficients from the event-study version of equation (1) equal zero. Standard errors clustered at the course level.

Table A4. Grade Effects by Alternative AI Exposure Measures

Outcome	Write + Code (dedup)	All tasks	Any task (0/1)	Writing only	Coding only
Share of A	0.13 (0.03)***	0.12 (0.03)***	0.04 (0.02)**	0.13 (0.03)***	0.03 (0.13)
Share of A-	-0.04 (0.01)***	-0.04 (0.01)***	-0.01 (0.01)**	-0.04 (0.02)***	0.01 (0.06)
Share of B+	-0.03 (0.01)***	-0.03 (0.01)***	-0.01 (0.01)*	-0.03 (0.01)***	-0.04 (0.04)
Share of B	-0.01 (0.01)	-0.01 (0.01)	-0.00 (0.01)	-0.02 (0.01)	0.08 (0.04)*
Share of B-	-0.02 (0.01)***	-0.02 (0.01)***	-0.01 (0.00)***	-0.02 (0.01)***	-0.02 (0.04)
Share of C	-0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)	-0.01 (0.04)
Share of D and F	-0.01 (0.01)	-0.01 (0.01)	-0.00 (0.00)	-0.00 (0.01)	-0.02 (0.02)
GPA	0.12 (0.05)**	0.12 (0.05)**	0.03 (0.02)	0.12 (0.05)**	0.06 (0.19)
GPA SD	-0.09 (0.04)**	-0.09 (0.04)**	-0.05 (0.02)***	-0.08 (0.04)**	-0.17 (0.14)
Share of Other	0.00 (0.01)	0.00 (0.01)	0.00 (0.00)	0.01 (0.01)	-0.04 (0.04)
Enrollment	19.85 (27.92)	8.89 (24.25)	-5.74 (16.10)	17.59 (28.71)	41.07 (68.80)
Observations	2,552	2,552	2,552	2,552	2,552
Courses	319	319	319	319	319

Notes: Each column reports pooled post-treatment estimates from equation (1) using a different AI exposure measure. Standard errors clustered at the course level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Table A5. Grade Effects by Minimum Enrollment Threshold

Outcome	n > 20 (main)	n > 15	n > 30	n > 40
Share of A	0.13 (0.03)***	0.12 (0.03)***	0.13 (0.04)***	0.14 (0.04)***
Share of A-	-0.04 (0.01)***	-0.03 (0.01)**	-0.04 (0.02)**	-0.04 (0.02)**
Share of B+	-0.03 (0.01)***	-0.03 (0.01)**	-0.03 (0.01)**	-0.03 (0.01)**
Share of B	-0.01 (0.01)	-0.02 (0.01)	-0.02 (0.01)	-0.03 (0.02)
Share of B-	-0.02 (0.01)***	-0.03 (0.01)***	-0.02 (0.01)***	-0.02 (0.01)***
Share of C	-0.01 (0.01)	-0.01 (0.01)	-0.02 (0.01)	-0.02 (0.01)*
Share of D and F	-0.01 (0.01)	-0.01 (0.01)	-0.00 (0.01)	-0.01 (0.01)
GPA	0.12 (0.05)**	0.12 (0.05)**	0.12 (0.05)**	0.13 (0.06)**
GPA SD	-0.09 (0.04)**	-0.07 (0.04)*	-0.08 (0.04)**	-0.08 (0.04)*
Share of Other	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	0.01 (0.01)
Enrollment	19.85 (27.92)	10.18 (22.55)	24.91 (32.36)	32.31 (38.31)
Observations	2,552	2,960	2,064	1,744
Courses	319	370	258	218

Notes: Each column reports pooled post-treatment estimates from equation (1), varying the enrollment threshold used to construct the balanced panel. Standard errors clustered at the course level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Table A6. Grade Effects by Pre-Period Sample

Outcome	Main (2018-2025)	Drop Fall 2020	2021-2025 only
Share of A	0.13 (0.03)***	0.13 (0.03)***	0.11 (0.03)***
Share of A-	-0.04 (0.01)***	-0.04 (0.02)***	-0.03 (0.02)*
Share of B+	-0.03 (0.01)***	-0.04 (0.01)***	-0.04 (0.01)***
Share of B	-0.01 (0.01)	-0.02 (0.01)	-0.02 (0.01)
Share of B-	-0.02 (0.01)***	-0.02 (0.01)***	-0.02 (0.01)**
Share of C	-0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)
Share of D and F	-0.01 (0.01)	-0.00 (0.01)	-0.00 (0.01)
GPA	0.12 (0.05)**	0.11 (0.05)**	0.10 (0.04)**
GPA SD	-0.09 (0.04)**	-0.08 (0.04)**	-0.09 (0.04)**
Share of Other	0.00 (0.01)	-0.00 (0.01)	0.00 (0.01)
Enrollment	19.85 (27.92)	18.05 (25.15)	-12.70 (26.73)
Observations	2,552	2,233	2,450
Courses	319	319	490

Notes: Each column reports pooled post-treatment estimates from equation (1), varying the pre-period sample. "Drop Fall 2020" excludes Fall 2020; "2021-2025 only" restricts the sample to 2021-2025. Standard errors clustered at the course level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Table A7. Grade Effects by Oral Presentation Task Exposure (Placebo)

Outcome	Presentation tasks (placebo)
Share of A	0.06 (0.09)
Share of A-	-0.04 (0.05)
Share of B+	0.00 (0.03)
Share of B	-0.07 (0.05)
Share of B-	-0.02 (0.02)
Share of C	0.01 (0.03)
Share of D and F	-0.00 (0.01)
GPA	0.03 (0.13)
GPA SD	-0.04 (0.11)
Share of Other	0.05 (0.03)*
Enrollment	-32.27 (67.27)
Observations	2,552
Courses	319

Notes: Pooled post-treatment estimates from equation (1), using the share of oral presentation tasks as the exposure measure. Standard errors clustered at the course level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Table A8. Grade Effects by AI Exposure Controlling for Instructor Stability

Outcome	+ Instructor stability
Share of A	0.13 (0.03)***
Share of A-	-0.04 (0.01)***
Share of B+	-0.03 (0.01)***
Share of B	-0.01 (0.01)
Share of B-	-0.02 (0.01)***
Share of C	-0.01 (0.01)
Share of D and F	-0.01 (0.01)
GPA	0.12 (0.05)**
GPA SD	-0.09 (0.04)**
Share of Other	0.00 (0.01)
Enrollment	19.79 (27.68)
Observations	2,552
Courses	319

Notes: Pooled post-treatment estimates from equation (1), adding instructor stability as a time-varying control. Standard errors clustered at the course level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Table A9. Grade Effects by AI Exposure with Homework-Weight Controls

Outcome	(1) Baseline	(2) + HW continuous	(3) + HW median split
Share of A	0.13 (0.03)***	0.11 (0.03)***	0.12 (0.03)***
Share of A-	-0.04 (0.01)***	-0.03 (0.02)**	-0.04 (0.01)**
Share of B+	-0.03 (0.01)***	-0.03 (0.01)**	-0.03 (0.01)**
Share of B	-0.01 (0.01)	-0.01 (0.01)	-0.02 (0.01)
Share of B-	-0.02 (0.01)***	-0.02 (0.01)***	-0.02 (0.01)***
Share of C	-0.01 (0.01)	-0.02 (0.01)	-0.01 (0.01)
Share of D and F	-0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)
GPA	0.12 (0.05)**	0.13 (0.05)***	0.12 (0.05)**
GPA SD	-0.09 (0.04)**	-0.10 (0.04)***	-0.09 (0.04)**
Share of Other	0.00 (0.01)	0.01 (0.01)	0.00 (0.01)
Enrollment	19.85 (27.92)	8.63 (23.39)	11.41 (25.92)
Observations	2,552	2,552	2,552
Courses	319	319	319

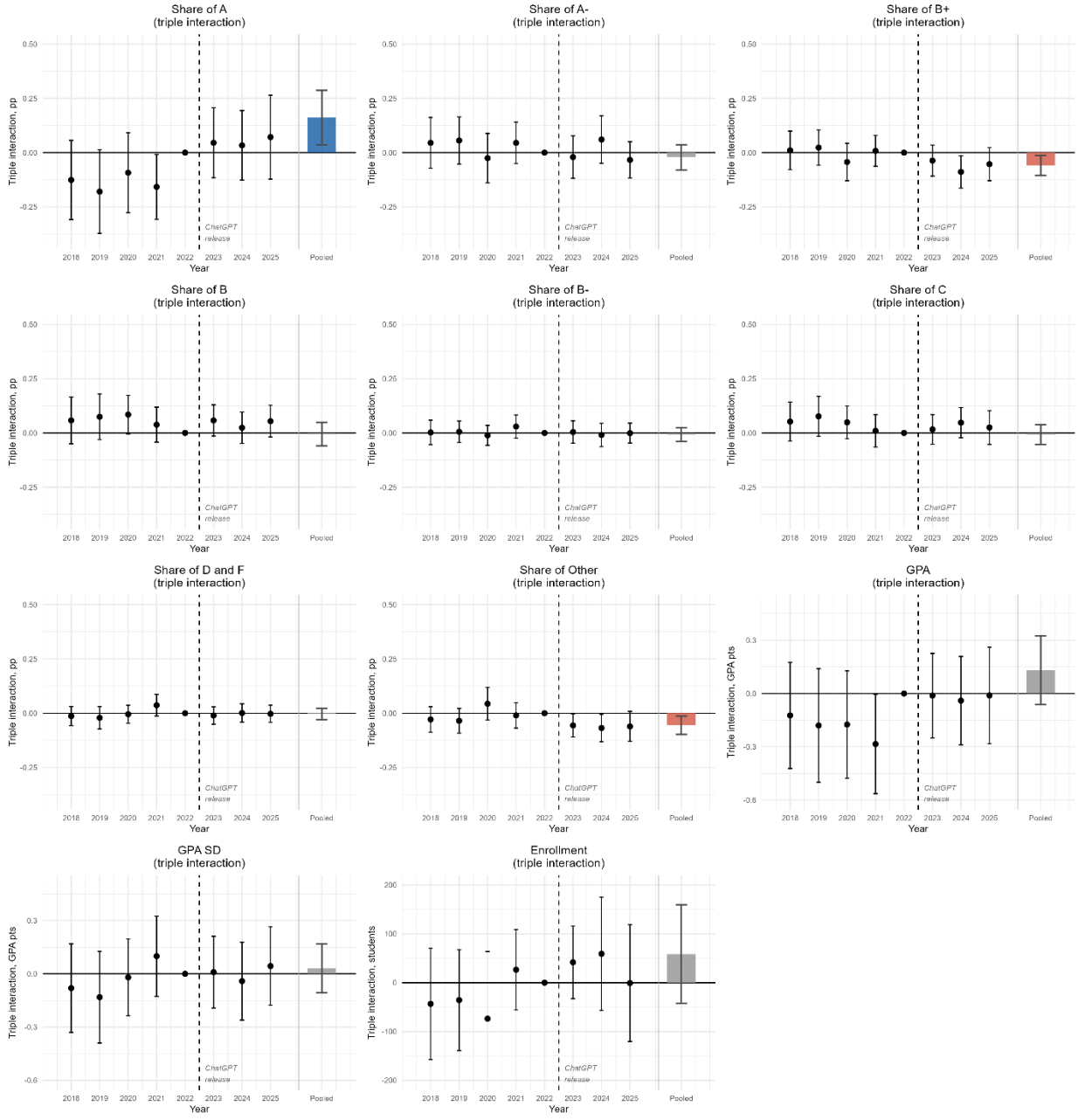
Notes: Each column reports the pooled post-treatment AI exposure estimate from equation (1). Columns (2) and (3) add post x homework-weight controls using continuous homework weight and the high-homework indicator, respectively. Standard errors clustered at the course level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Table A10. Grade Effects by AI Exposure and Homework Weight (Cumulative Thresholds)

Outcome	Post x Exposure (low HW)	Post x Exposure x High HW
At least A	0.017 (0.046)	0.161 (0.064)**
At least A-	-0.005 (0.046)	0.139 (0.061)**
At least B+	0.004 (0.044)	0.080 (0.058)
At least B	-0.008 (0.037)	0.074 (0.047)
At least B-	-0.025 (0.031)	0.067 (0.040)*
At least C-	-0.035 (0.024)	0.059 (0.028)**
Observations	2,552	2,552
Courses	319	319

Notes: Pooled post-treatment estimates from equation (2) that are shown in Figure 2. Standard errors clustered at the course level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Figure A2. Grade Effects by AI Exposure and Homework Weight



Notes: Each panel shows event-study estimates from equation (2) for the share of a given grade and summary measures reported in Table 3. The dashed vertical line marks ChatGPT's release in November 2022. The bar to the right of each panel shows the pooled triple interaction estimate from the same specification. Bars are colored blue if the pooled estimate is positive and significant at the 5% level, red if negative and significant, and gray otherwise. Whiskers show 95% confidence intervals. Standard errors clustered at the course level.

Table A11. Grade Effects by AI Exposure and Homework Weight (Department-Clustered Standard Errors)

Outcome	Post x Exposure (low HW)	Post x Exposure x High HW
Share of A	0.017 (0.054)	0.161 (0.080)**
Share of A-	-0.022 (0.025)	-0.022 (0.041)
Share of B+	0.009 (0.017)	-0.059 (0.022)***
Share of B	-0.012 (0.020)	-0.006 (0.026)
Share of B-	-0.017 (0.015)	-0.007 (0.020)
Share of C	-0.009 (0.020)	-0.008 (0.023)
Share of D and F	-0.004 (0.010)	-0.004 (0.013)
Share of Other	0.039 (0.021)*	-0.056 (0.025)**
GPA	0.044 (0.087)	0.132 (0.101)
GPA SD	-0.106 (0.047)**	0.031 (0.062)
Enrollment	-24.481 (33.060)	58.605 (47.505)
Observations	2,552	2,552
Courses	319	319

Notes: Pooled post-treatment estimates from equation (2), using the same specification as Table 3. Standard errors clustered at the department level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Table A12. Pre-Trend Tests by AI Exposure and Homework Weight

Outcome	Chi-square	p-value
Share of A	5.665	0.226
Share of A-	3.726	0.444
Share of B+	2.554	0.635
Share of B	3.737	0.443
Share of B-	2.992	0.559
Share of C	3.726	0.444
Share of D and F	5.999	0.199
Share of Other	5.067	0.281
GPA	4.062	0.398
GPA SD	3.159	0.532
Enrollment	4.419	0.352
Observations	2,552	
Courses	319	

Notes: Joint Wald tests of whether the 2018-2021 year-by-exposure-by-high-homework coefficients from the event-study version of equation (2) equal zero. Standard errors clustered at the course level.

Table A13. Grade Effects by AI Exposure and Homework Weight Across Specifications

Outcome	Main (median, imputed)	Observed only	Mean split	n > 30	n > 40
Share of A	0.16 (0.06)**	0.15 (0.07)**	0.15 (0.07)**	0.15 (0.07)**	0.15 (0.07)**
Share of A-	-0.02 (0.03)	-0.02 (0.03)	-0.02 (0.03)	-0.05 (0.03)	-0.05 (0.04)
Share of B+	-0.06 (0.02)**	-0.05 (0.02)**	-0.05 (0.02)**	-0.06 (0.02)**	-0.04 (0.03)
Share of B	-0.01 (0.03)	-0.00 (0.03)	-0.02 (0.03)	0.01 (0.03)	-0.00 (0.03)
Share of B-	-0.01 (0.02)	-0.01 (0.02)	-0.02 (0.02)	-0.00 (0.02)	-0.00 (0.02)
Share of C	-0.01 (0.02)	-0.01 (0.02)	-0.00 (0.02)	-0.01 (0.02)	-0.00 (0.03)
Share of D and F	-0.00 (0.01)	-0.00 (0.01)	-0.00 (0.01)	-0.01 (0.01)	0.00 (0.02)
GPA	0.13 (0.10)	0.13 (0.10)	0.13 (0.10)	0.12 (0.11)	0.10 (0.12)
GPA SD	0.03 (0.07)	0.01 (0.07)	0.04 (0.07)	0.04 (0.08)	0.06 (0.09)
Share of Other	-0.06 (0.02)***	-0.05 (0.02)**	-0.04 (0.02)**	-0.04 (0.02)**	-0.06 (0.02)**
Enrollment	58.60 (51.55)	56.75 (53.37)	67.99 (51.18)	64.40 (61.10)	73.22 (74.35)
Observations	2,552	2,488	2,552	2,064	1,744
Courses	319	311	319	258	218

Notes: Each column reports the pooled post-treatment triple interaction estimate from equation (2), varying the homework-weight measure or enrollment threshold. “Observed only” excludes 8 courses with imputed homework weights; “Mean split” defines high homework using the mean rather than median. Standard errors clustered at the course level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.