

From A-Minus to A: Generative AI and Grade Compression at the Top

Reza Tavakoli Jaghargh

Department of Economics, Virginia Tech, rtavakoli@vt.edu

Draft V4, October 2026

Abstract

Grades help employers and graduate programs tell students apart, but they have been rising for decades and are now close to the top of the scale. This paper studies whether generative AI pushed this trend further, using data from the University of Texas at Austin. The data are section-level grade distributions and section-level syllabi, for 2015 to 2025, and Texas law requires public universities to post the syllabi online. Every graded component in each syllabus is coded by a large language model for its weight, its task type and whether it is proctored. Because the sample ends in 2025, the results describe the effect of the first generation of generative AI tools, which students used from 2023 to 2025. The main measure, AI exposure, is the share of the course grade that comes from work that AI can do and that students complete without a proctor. After the release of ChatGPT, courses with higher AI exposure saw more A grades and fewer A- and B+ grades, while low grades did not change. A move from the 25th to the 75th percentile of AI exposure raised the share of A by 4.5 percentage points, about 10 percent of its mean before 2023 (44.6 percent), and lowered the shares of A- and B+ by 1.8 and 1.1 percentage points, about 11 percent of their means. Instructors responded quickly by adding AI policies to their syllabi, yet they barely moved grade weight away from tasks that AI can do, and they did not proctor more of the graded work.

JEL codes: I21, I23, J24, O33.

Keywords: generative AI, college grades, grade compression, course syllabi, large language models.

1 Introduction

Undergraduate grades have risen substantially in the United States over the past two decades (Rojstaczer and Healy 2012), raising questions about their ability to distinguish students' academic achievement. Average undergraduate GPA increased from 2.89 in 2000 to 3.15 in 2020.[1] At selective universities, the growing concentration of grades at the top is particularly striking: the share of A grades at Harvard increased from 24 percent in 2005 to 60 percent in 2024-25, while the share of grades in the A range at Yale rose from 67 percent in 2010-11 to 79 percent in 2022-23 (Harvard Office of Undergraduate Education 2025; Fair 2023). Although rising grades can reflect improvements in learning, evidence that students receive higher course grades despite identical examination performance indicates that changes in grading standards also contribute (Denning et al. 2022). These developments have potential labor-market consequences because employers use GPA to evaluate graduates before observing their productivity (Spence 1973; Hansen, Hvidman and Sievertsen 2024). Rising grades may therefore weaken the signaling value of GPA if they reduce its ability to distinguish among students with different levels of proficiency. Because GPA has an upper bound, this can occur through bunching at the top of the grading scale, whether the rise is driven by improvements in proficiency or by more generous grading, so that students with different skills are pooled near the ceiling and employers can draw fewer distinctions from academic records (Chan, Hao and Suen 2007; Ostrovsky and Schwarz 2010). The main concern is thus not that average grades are higher, but that grades carry less information about students, especially at the top of the grade distribution.

The emergence of generative artificial intelligence (GenAI) introduces a new potential source of change in both the level and distribution of students' GPAs. Within two years of the release of ChatGPT, more than 80 percent of students at a selective US college reported using AI for their coursework (Contractor and Reyes 2026). By assisting with writing, coding, and problem-solving, GenAI may enable students to obtain higher grades through improved learning (a learning effect), through assistance in completing assessed work (a substitution effect), or both, so its effects on the distribution of grades are ambiguous. If students below the upper end of the GPA distribution experience larger gains, GenAI could narrow grade differences and increase the concentration of students near the ceiling; differences in access and effective use could also

alter students' relative positions. Such changes would weaken GPA's signaling value only insofar as they obscure differences in skills relevant to workplace productivity. This possibility is consistent with experimental evidence that employers place less weight on grades when grading schemes distinguish less finely between candidates (Pu, Abel and Carpenter 2026). Understanding GenAI's consequences therefore requires examining more than its effect on average grades. This paper investigates how GenAI has affected students' GPAs and their distribution, with particular attention to concentration near the grade ceiling and changes in the upper tail, and it also examines how instructors responded to this shock through the rules on AI use and the grading of their courses.

The analysis rests on two public sources: section-level grade distributions, which report the number of students who received each letter grade in each course section, and section-level syllabi, which a Texas law has required public universities to post online for every undergraduate course section since 2010. The syllabi therefore give a record of how each section was graded, both before and after the release of ChatGPT in November 2022, the first GenAI tool in wide use among students. In each syllabus, every graded component was coded by a large language model for its weight in the final grade, its task type and whether students complete it under a proctor, and the proctoring codes agree closely with a hand coding of 500 syllabi by a research assistant. These codes are combined into one measure, AI exposure, which is the share of the course grade that comes from work that AI can do at a passing level and that students complete without a proctor. Exposure is measured from the 2022 syllabi, which were written before ChatGPT was released, and the main sample is a balanced panel of 634 undergraduate courses observed from 2018 to 2025, with about 1.25 million letter grades, and the main estimates use its 493 courses offered in every fall. A difference-in-differences design with a continuous treatment, together with an event study, compares changes in grades after November 2022 across courses with differing AI exposure. The identifying assumption is that, absent ChatGPT, grade distributions would have followed parallel trends across exposure levels. Measuring exposure before the release avoids defining treatment using instructors' subsequent responses, although differential changes associated with the recovery from pandemic teaching remain a potential concern. To assess sensitivity to the exposure baseline, the analysis also constructs exposure from 2019 syllabi, before the pandemic altered assessment practices.

After the release of ChatGPT, courses with higher AI exposure saw more A grades and fewer A- and B+ grades, while the shares of low grades did not change. A move from the 25th to the 75th percentile of AI exposure, which equals about 45 percentage points of the course grade, raised the share of A by 4.5 percentage points, about 10 percent of its mean before 2023 (44.6 percent; Table 10). The same move lowered the shares of A- and B+ by 1.8 and 1.1 percentage points, which are about 11 percent of their means, and the shares of F and of D or F did not change. Estimates for each letter grade show that the rise in the share of A was matched almost entirely by falls in the shares of A-, B+ and B- (Figure 6). Mean grade points rose by only 0.035 grade points for the same move (Table 10), indicating that the main effect of AI exposure is a move from just below the top grade to the top grade rather than a general rise in grades. The shift from A- and B+ grades to A grades remains when exposure is constructed from 2019 rather than 2022 syllabi, although the rise in the share of A is smaller (2.1 percentage points; Online Appendix Table B2), so the central finding does not depend on a baseline potentially shaped by pandemic assessment practices. As a result, the grade distribution became more compressed at the top, where grades matter most for telling the strongest students apart.

Instructors could have limited these effects by changing their courses, and this response is studied with a second panel of 529 instructors whose syllabi are observed in every fall or every spring from 2018 to 2025. Instructors responded mainly through rules, since the share of syllabi with a policy on AI use rose from 2 percent in 2022 to 50 percent in 2025, and policies were more common in courses where AI could do more of the graded work (Figures 9 and 10). The design of courses changed much less, because within the same instructor and course, the share of the grade from task types that AI can do fell by only 1.7 percentage points from 2022 to 2025, about 4 percent of its 2022 mean (Figure 12). Changes in proctoring were also limited: the unproctored share fell after 2022 only in courses whose unproctored share in 2022 was still more than 5 percentage points above its 2019 level (Figure 14). This pattern is consistent with a reversal of the pandemic-era move toward take-home work, although it does not separately identify the roles of pandemic recovery and responses to AI. Taken together, the syllabus evidence suggests that the 2022 exposure measure continued to capture the broad structure of assessed work after ChatGPT's release. Instructors may nevertheless have responded in ways that leave no trace in the syllabus, such as using AI detectors or making assignments harder for AI.

The section-level data cannot separate the learning effect from the substitution effect, although field evidence shows that substitution is real: high school students with unrestricted access to GPT-4 did better on practice problems but worse on a later exam taken without AI (Bastani et al. 2025), and secondary school students in China who adopted AI scored higher on homework but lower on exams (Strömberg, Lei and Wu 2026). Both channels, however, compress grades at the top, because the 4.0 ceiling hides the gains of the strongest students even when every student learns more.

The estimates in this paper describe the first generation of GenAI tools, which students used from 2023 to 2025, and these tools have improved quickly since then. On Humanity's Last Exam, a set of 2,500 expert-level questions, GPT-4o answered 2.7 percent of the questions correctly in November 2024, while the best OpenAI model answered 53.6 percent in September 2026 (Center for AI Safety and Scale AI 2026),[2] so the effects of the tools that students use today may be larger. The data also report grades only for whole sections, so they cannot show whether weaker students who used AI on take-home work then did worse on proctored exams in the same course, which would offset part of the rise in grades.

This paper makes three contributions, and the first is a new measure of course-level AI exposure that jointly accounts for AI capability and proctoring, which contributes to the literature on the measurement of AI exposure (Felten, Raj and Seamans 2021; Eloundou et al. 2024). The closest study, Chirikov (2026), uses syllabi and grade distributions from a large university in Texas, measures exposure as the share of required tasks that are writing or coding, and finds that full exposure raises the share of A by 13 percentage points. Hausman, Rigbi and Weisburd (2025) instead compare courses graded at home with courses graded in person at an Israeli university, and they find gains that are largest at the bottom of the distribution. The measure developed here integrates these two dimensions at the level of each graded component: whether AI can perform the task at a passing level and whether students complete it without a proctor. Weighting these components by their contribution to the final grade captures both technological capability and the assessment conditions under which students can use it. To the author's knowledge, this is the first measure of course-level AI exposure to combine these dimensions in this way. Its construction from public syllabi also provides an approach that can be applied to other universities.

Second, the paper adds evidence to the small but growing literature on GenAI and student performance (Bastani et al. 2025; Hausman, Rigbi and Weisburd 2025; Chirikov 2026; Strömberg, Lei and Wu 2026; Zumel Dumlao et al. 2026). Using a large panel of undergraduate courses, it documents changes concentrated at the upper end of the grade distribution, with increases in A grades accompanied by declines in A- and B+ grades and little change in low grades. Third, to the author's knowledge, this is the first study to examine instructors' responses to GenAI systematically using a longitudinal panel of syllabi. By tracking both stated AI policies and assessment design, the analysis distinguishes changes in course rules from changes in the work that determines students' grades. It shows that policies on AI use spread rapidly while assessment design changed relatively little.

The rest of the paper is organized as follows: Section 2 describes the setting and the data, Section 3 explains how AI exposure is measured from the syllabi, and Section 4 presents the empirical strategy. Section 5 reports the main results, Section 6 studies how instructors responded to GenAI, Section 7 summarizes the robustness checks, which are reported in the Online Appendix, and Section 8 concludes.

2 Setting and Data

2.1 Institutional Setting

This paper studies changes in grades following the release of ChatGPT using linked course records from the University of Texas at Austin (UT Austin). The data combine section-level grade distributions, section-level syllabi, and instructor curriculum vitae (CVs), which are linked to each other through the course section. Grade distributions measure academic outcomes, while syllabi describe assessment components, their weights in the final grade, and whether assessments are completed in class or at home. Instructor CVs provide information on educational qualifications and employment histories, so that, together, these sources make it possible to examine grades alongside assessment design and instructor characteristics.

The availability of syllabi and instructor CVs reflects a statewide disclosure requirement that has applied to public universities in Texas since fall 2010. Texas House Bill 2504, codified in Section 51.974 of the Texas Education Code, requires public institutions of higher education, other than medical and dental units, to publish syllabi for undergraduate classroom courses

offered for credit and CVs for their regular instructors. Syllabi must include a brief description of each major assignment and examination, and the documents must be publicly accessible without registration, posted no later than seven days after the first day of classes, and retained online for at least two years.[3] Mandatory disclosure reduces reliance on instructors' voluntary decisions to share course materials, although it does not ensure complete coverage. UT Austin also publishes the grade distribution of each course section through a public dashboard maintained by its office of EVPP Analytics and Reporting (UT Austin, EVPP Analytics and Reporting n.d.).[4]

UT Austin offers a large and academically diverse setting, since in fall 2025 the university enrolled approximately 55,000 students, including 44,314 undergraduates (UT Austin 2025a). Although syllabi are available from 2010, the document collection assembled for this paper covers fall and spring semesters from 2015 through 2025 and contains 131,015 syllabi and 129,926 section-linked CV records. These records span 131,026 course sections, approximately 7,100 courses, and 226 subject areas, so the results do not rest on a small number of fields.[5]

The observation period supports comparisons within the same courses over many semesters before and after the release of ChatGPT, with eight years before the release (2015-2022) and three years after it (2023-2025), and the main estimation sample uses the years from 2018 to 2025. ChatGPT became publicly available on November 30, 2022, near the end of the fall semester, so spring 2023 is the first full semester following its release.

UT Austin's guidance leaves decisions about permitted instructional uses of generative AI to individual instructors, subject to university policies. Guidance updated in November 2023 recommends that instructors communicate these decisions in their syllabi and notes that unauthorized uses may constitute academic dishonesty (UT Austin 2026). As a result, the rules on AI use were set course by course, which provides a setting in which to examine how instructors responded to generative AI.

Although the sample is restricted to courses offered consistently over time, and UT Austin is a selective public research university, the courses in the sample cover a wide range of fields. The results are most informative for other large public universities and may apply less directly to small colleges or to less selective institutions.

Figure 1 reports the number of available syllabi in each fall and spring semester from 2015 through 2025. The grey area marks the period of pandemic-related instructional disruption, from spring 2020 to spring 2021, and the blue area marks the period after the release of ChatGPT. UT Austin moved instruction online during spring 2020 and planned a broad return to in-person instruction for fall 2021 (Texas Tribune 2020; Reporting Texas 2021). The number of syllabi ranges from approximately 5,400 to 6,800 per semester, with no decline during the pandemic or after the release of ChatGPT, and the rise in 2024 and 2025 follows the growth in enrollment.

Documents are linked to grade distributions using the year, the semester, and the unique section number of the university. Approximately 76 percent of sections in the grade data have a matching syllabus, while 71 percent of downloaded syllabi have matching grade data, and two differences in the coverage of the two sources explain most of this gap.

First, the statutory syllabus requirement applies to undergraduate classroom courses, whereas the grade data also include graduate and professional courses. Graduate and professional sections account for more than half of the sections in the grade data without a matching syllabus. Among sections in the grade data, syllabus coverage is 22 percent for graduate and professional courses and 88 percent for undergraduate courses, and the undergraduate rate rises to 92 percent when sections are weighted by enrollment.

Second, to protect student privacy, the university does not report grades for undergraduate sections with fewer than 10 students, so the syllabi of small undergraduate sections have no matching grade data. The only sections below this size with course numbers resembling undergraduate numbers are 135 sections of law, pharmacy, and medical courses, which are classified as graduate or professional courses in the analysis. For the same reason, the share of downloaded syllabi with matching grade data is lower for upper-division courses, which are often small, than for lower-division courses (67 percent against 80 percent).

The share of undergraduate sections with a syllabus changes only slightly over time, from about 86 percent in 2015 to 2019 to about 90 percent in 2023 to 2025,[6] so coverage changed little in response to COVID-19 or to ChatGPT. These aggregate rates do not show, however, that the sections without a syllabus or without grades are similar to the others. Section 2.2 describes the three data sources in more detail, and Section 2.3 details the construction of the two balanced panels used in the analysis.

2.2 Data Sources

This paper combines three sources: section-level grade distributions, section-level syllabi, and instructor CVs. The grade data are structured records that report the number of students receiving each letter grade in a section, whereas syllabi and CVs vary in format: grading schemes may appear in tables, lists, or prose, and CVs differ in how they report academic appointments. Large language models (LLMs) are used to extract structured variables from both document types, and the coding and validation of the syllabus measures are described in Section 3, while Online Appendix A documents the models and procedures used to code instructor CVs and Online Appendix E prints the prompts.

Grades. The grade data report the number of students receiving each letter grade by course section and semester, together with the course prefix, number, title, and unique section number. Letter grades are converted to grade points using the official scale of UT Austin, ranging from 4.00 for an A to 0.00 for an F. Symbols excluded from the university's GPA calculation, including credit, no credit, Q (drop), and W (withdrawal), are also excluded from the grade measures in this paper (UT Austin 2025b). A+ grades appear only in School of Law courses, which are professional courses and are excluded from the analysis.

Table 1 summarizes undergraduate grade distributions for fall and spring semesters from 2015 through 2025, covering 100,806 sections and approximately 3.4 million letter-grade observations. Annual grade shares and mean grade points are calculated by pooling letter-grade counts across sections, which gives each recorded letter grade equal weight. Grade dispersion is measured by the standard deviation of grade points within each section, and the table reports the mean of these within-section standard deviations, weighted by the number of letter grades in each section.

Grades increased well before the release of ChatGPT: between 2015 and 2019, mean grade points rose from 3.24 to 3.35, while the share of A grades increased from 34.3 to 40.6 percent. Both measures rose sharply in 2020, reaching 3.61 and 57.2 percent, before declining to 3.44 and 49.8 percent in 2022. Comparisons involving 2020 require caution because the share of undergraduate enrollments receiving symbols other than letter grades also increased, from 6-7 percent in 2015-2019 to 14 percent in 2020. Changes in observed grades in that year may therefore partly reflect changes in which enrollments received letter grades.

Grades rose again after 2022, and by 2025 mean grade points had reached 3.55, the share of A grades was 56.1 percent, and the share receiving A- or above was 70.1 percent. Between 2022 and 2025, the mean within-section SD declined from 0.70 to 0.61, while the share of F grades fell from 1.8 to 1.1 percent. Grades therefore became both higher and more compressed in the years after the release of ChatGPT, although these aggregate patterns may also reflect changes in the composition of sections and enrollments.

Syllabi and instructor CVs. Syllabi and CVs were collected with a web scraper from the course documents website of UT Austin, which provides section-specific links to these documents for fall and spring semesters. Summer terms are excluded because they differ in duration, course offerings, and student composition, and a corresponding instructor CV is available for approximately 99 percent of sections with a syllabus.

The retrieved CVs reflect the versions available at the time of collection rather than the versions that were current in each past semester. Earlier positions of instructors are therefore reconstructed from the appointment titles and dates reported in these CVs, so the accuracy of these historical measures depends on the completeness and precision of the reported employment histories.

Table 2 summarizes the syllabus collection by year: annual counts range from 11,064 to 13,314 syllabi, associated with 2,920 to 3,303 distinct instructor names and 167 to 202 subject areas. Upper-division courses account for approximately half of the syllabi each year, lower-division courses for approximately 40 to 45 percent, and graduate or professional courses for the remainder, and Section 2.3 describes the sample restrictions and the construction of the balanced panels.

2.3 Sample Construction

The empirical analysis in this paper relies on two balanced panels built from the documents described above: a panel of consistent courses and a panel of consistent instructors. In a balanced panel, the same units are observed in every year, both before and after the release of ChatGPT, so that changes over time are not driven by a changing set of courses or instructors. The set of courses or instructors is fixed, but enrollment and the students in each course can still change. The panel of consistent courses is used to study grades (Sections 4 and 5), and the panel of

consistent instructors is used to study how instructors changed their syllabi (Section 6). Further restrictions that apply to a single design, such as the base year of the exposure measure or the semester that is used, are described together with that design.

2.3.1 Consistent Courses

A course is defined by its prefix, number, and title, because some course numbers at UT Austin are used for several different courses. Courses whose title covers changing subjects, such as special topics, independent study, internships, and theses, are dropped, because their content changes from year to year. A course is consistent if it is offered in every fall, or in every spring, of the time window. A course that qualifies in both semesters contributes sections from both; otherwise, only the sections of the qualifying semester are used. Two time windows are used: 2015 to 2025 and 2018 to 2025.

Table 3 compares the panels with all sections that have a downloaded syllabus from 2015 to 2025. The panel for 2018 to 2025 contains 1,027 courses and 38,081 sections. Of these courses, 774 are offered in every fall, 641 in every spring, and 388 in both semesters. The panel is much smaller than the full set of 7,081 courses, which are counted by prefix and number, because most courses at UT Austin are not offered in every year and because topics courses are dropped.

The courses that remain are larger and more often introductory: the average section has about 44 students, compared with 37.5 students for all sections, and 56 percent of sections are lower-division, compared with 42 percent for all sections, while graduate or professional courses make up only 1.8 percent of sections. Grades in the panel are close to those of all sections, with mean grade points of 3.40 compared with 3.41 and a share of A grades of 46.4 percent compared with 45.9 percent.

Table 4 describes the instructors who taught these courses in 2025, based on the current CV of each instructor. The panel for 2015 to 2025 has 1,161 instructors in 2025 and the panel for 2018 to 2025 has 1,524. In both panels, about 60 percent of instructors hold a doctorate, and the average instructor has been at UT Austin for 12.6 years and received the highest degree 16 years earlier. About one third of instructors are tenured or tenure-track faculty, a little more than one third are teaching faculty, such as lecturers and professors of practice, and 11 to 12 percent are

graduate student instructors. The role of about 16 percent of instructors cannot be classified from the CV.

2.3.2 Consistent Instructors

An instructor is consistent if the instructor has at least one syllabus in every fall from 2018 to 2025 (the fall track), or in every spring from 2018 to 2025 (the spring track), in any course. Unlike the course panel, this panel does not require the instructor to teach the same course in every year. This panel measures changes in syllabi within the same instructor, although the courses that an instructor teaches may change from year to year.

Table 5 describes the panel, which contains 529 instructors: 403 on the fall track and 335 on the spring track, of whom 209 are on both tracks. Together, these instructors taught 20,585 sections of 2,150 courses in 152 subject areas between 2018 and 2025.

On average, each instructor-track pair has about 3.5 sections of 2 courses per year, and in 44 percent of these pairs the instructor taught the same course in all eight years of the track. About 64 percent of their sections belong to a course in the panel of consistent courses for 2018 to 2025, so the two panels overlap but are not the same. The mix of course levels and grades in the instructor panel is close to that of all sections: 43 percent of sections are lower-division and 52 percent upper-division, and mean grade points are 3.44, with 48.2 percent A grades.

3 Measuring AI Exposure

This section explains how course syllabi are turned into a measure of AI exposure, which is built for each graded component of a course, such as an essay, a quiz, or a final exam, and then added up to the course level. Section 3.1 explains the logic of the measure and how each of its two dimensions is coded. Section 3.2 describes how the syllabi were coded with large language models (LLMs) and how this coding was checked against human coders. Section 3.3 describes the three course-level indices used in the rest of the paper.

The measure improves on the one in Chirikov (2026), the closest study, in three ways. Chirikov measures exposure as the share of writing and coding tasks among the distinct tasks listed in a syllabus. The measure in this paper also counts quizzes and multiple-choice tests, which AI could answer at a passing level by 2024-25. It weights each component by its share of the final grade, rather than counting tasks, and it also asks, for each component, whether the work is done

without a proctor. As a result, the measure captures the part of the grade that AI can actually move: work that AI can do, that carries weight in the grade, and that students complete without being seen.

3.1 Logic of the Measure

Generative AI can change a course grade only when two conditions hold for the graded work. First, AI must be able to do the task well enough to pass, and second, the student must be able to use AI without being seen, which means the work is completed without a proctor. A take-home essay meets both conditions, an essay written in a proctored exam room meets only the first, and an attendance grade meets neither. AI exposure is therefore measured along two dimensions in this paper: AI capability, which asks whether AI can do the task, and the unproctored share, which asks whether the task is completed without a proctor.

In this paper, both dimensions are measured for each graded component j of course c , where w_j is the weight of component j in the final grade. $Alcap_j$ equals one if the component belongs to a task type that AI can do at a passing level, and zero otherwise. $Unproc_j$ equals one only if the component is truly unproctored, meaning that the student completes it outside the view of the instructor or a proctor, as in a take-home essay or an online quiz without a proctor, and it equals zero otherwise. Work done in class is proctored, including a class presentation or an in-class discussion, and a component is also coded as zero when the syllabus does not say how it is completed. AI exposure of course c is

$$E_c = \frac{\sum_j w_j \times Alcap_j \times Unproc_j}{\sum_j w_j} \quad (1)$$

Because $Alcap_j$ and $Unproc_j$ are both binary, E_c has a simple meaning: it is the share of the course grade that comes from work that AI can do and that is completed without a proctor. The two dimensions are multiplied for each component before the components are added up, not as two course averages, and this order matters. For example, a course may have a proctored essay exam and an unproctored attendance grade. The product of its two course averages is positive, but its exposure is zero, because no single component meets both conditions.

Two additional indices summarize each dimension separately: the AI capability index is the grade-weighted share of components classified as within AI's capabilities, regardless of supervision. The unproctored share is the grade-weighted share classified as unproctored,

regardless of task type. Sections 4 and 5 compare AI capability with AI exposure to distinguish task composition from assessment conditions.

3.1.1 AI Capability by Task Type

Each graded component is assigned by the LLM to one of nine task types, listed in Table 6, but the LLM does not decide which types count as AI-doable. That decision comes from a fixed rule for $AIcap_j$, set before estimation, so the measure does not depend on which model reads the syllabus. AI capability has a narrow meaning in this paper, since it asks only whether AI could produce work that passes a typical undergraduate assignment of a given type. It does not ask whether AI is useful for the task in general, which is the question behind measures of occupational exposure such as Eloundou et al. (2024).

AI models improve quickly, and their abilities change from year to year and from task to task. AI capability is therefore coded for one base year and one model: the 2024-25 academic year, near the middle of the post-ChatGPT period studied in this paper (spring 2023 to fall 2025), and GPT-4o, the most capable model that ChatGPT offered for free in that year (OpenAI 2024b). A broad literature finds that models of this kind perform well on three kinds of graded work: written prose, code, and objective items such as multiple-choice and short-answer questions. Eloundou et al. (2024) find that programming and writing skills are the ones most strongly linked to exposure to large language models, and Chirikov (2026) treats writing and coding as the tasks where AI is strongest when he builds his exposure measure. Scarfe et al. (2024) find that essays written by GPT-4 and submitted into a real university exam system received higher grades than most student work,[7] and on the multiple-choice Test of Understanding in College Economics, ChatGPT scored in the 91st percentile in microeconomics and the 99th percentile in macroeconomics relative to students at the end of a principles course (Geerling et al. 2023).

Public benchmarks are used to check this grouping, under a rule that leaves little room to choose benchmarks that favor it. For each task type, the benchmark is taken from the set that OpenAI reported for GPT-4o in its simple-evals repository (OpenAI 2024a, 2024b), and a task type is coded as AI-doable ($AIcap = 1$) if GPT-4o scored at least 70 percent, a passing grade in most US courses. GPT-4o answered 88.7 percent of the questions in MMLU, a set of multiple-choice questions in 57 college subjects (Hendrycks et al. 2021a), and solved 90.2 percent of the programming tasks in HumanEval (Chen et al. 2021; OpenAI 2024a).[8] The one departure from

this rule concerns quantitative problem sets, because the reported benchmark, MATH, consists of short competition problems on which GPT-4o scored 75.9 percent (Hendrycks et al. 2021b), so the closer benchmark U-MATH, a set of problems from US university math courses on which GPT-4o solved only about half, is used instead (Chernyshev et al. 2024).[9] Written prose has no benchmark with a pass rate in this set, so its coding rests on the course-grading evidence above. Three types are therefore coded as AI-doable, namely written prose, objective items, and code or data (Table 6), and all other types are coded as zero. Oral work, studio and lab work, and participation are zero by construction, because the student produces the graded work in person, and components with no clear task, such as "Exam 2, 20 percent", are also set to zero. Table A1 in the Online Appendix lists every benchmark that OpenAI reported for GPT-4o, the task type to which each benchmark maps, and the resulting code.

As a robustness check, $Alcap_j$ is also replaced by the benchmark score of its task type instead of the binary value, which is 0.887 for objective items, 0.83 for written prose, 0.902 for code or data and 0.78 for problem sets. Components with no clear task receive either 0.873, the mean score of the three AI-doable types, or zero, and the results are reported in Online Appendix B.

3.1.2 Unproctored Work

The second dimension asks whether a student can use AI on a graded component without being seen. For each component, the LLM records how the work is completed, using one of seven categories: in person, online and live, online with a proctor, online and self-paced, take-home, not applicable, or unknown, and these categories are then mapped to $Unproc_j$. Work done in person, or during a live online session that the instructor watches, is coded as proctored ($Unproc_j = 0$). Other online work is coded as proctored only when the syllabus names a proctoring tool for that specific component, such as Proctorio, Honorlock, Respondus LockDown Browser, or ProctorU, or requires the webcam to be on. Take-home work and other online work, such as a problem set or an online quiz, are coded as unproctored ($Unproc_j = 1$). An online quiz or exam counts as unproctored even when it is timed or closed-book, because a time limit is not supervision. When the syllabus does not say how a component is completed, the component is coded as proctored ($Unproc_j = 0$).

The coding is strict on purpose, since a component counts as unproctored only when the syllabus makes it clear that the work is done without supervision. A class presentation, an in-class quiz, or

an exam with no stated format therefore counts as proctored. Because of this rule, the unproctored share is a lower bound on the share of the grade that students can complete with AI. Online Appendix B reports results in which components whose proctoring is indeterminate, because one weight covers both proctored and unproctored work, count as half unproctored ($Unproc_j = 0.5$), and the estimates barely change.

The unproctored share of course c is the grade-weighted mean of $Unproc_j$:

$$U_c = \frac{\sum_j w_j \times Unproc_j}{\sum_j w_j} \quad (2)$$

It is equal to zero for a course graded only on proctored work, and to one for a course graded only on take-home or unproctored online work, and the weights w_j come from the grading scheme in the syllabus. Section 3.2 explains how the weights and the format of each component are read from the syllabus, and Section 3.3 describes how the unproctored share is distributed across courses.

3.2 Coding Syllabi with LLMs

Because syllabi follow no common format, simple text tools such as keyword searches or regular expressions cannot code them reliably: the grading scheme may appear as a table, a list, or a paragraph, and the format of each component is often described in another part of the document. LLMs can read a whole syllabus in the way a person would, so the syllabi were coded by LLMs, with the same coding rules in every run. For each syllabus, the model returns one structured record that lists every graded component with its name, its weight, how it is completed (one of the seven categories in Section 3.1.2), and its task type (Table 6), together with the text in the syllabus that supports the task type. The model only reads and classifies, and the three measures are computed afterwards by the same code for every model, while the output format and all settings are given in Online Appendix A, and the full prompt is printed in Online Appendix E.

3.2.1 Model Choice

Two models were used, one for each panel: the course panel, which gives the main measure, was coded with Gemini 3.7 Flash from Google, one of the strongest frontier models available when the syllabi were coded. The instructor panel in Section 6 needed about 12,300 more syllabi to be coded, so it was coded with GLM-5.3, one of the strongest open-weight models, which runs at no

cost on the Advanced Research Computing cluster of Virginia Tech and which other researchers can run again.

The two models were chosen with a public benchmark of long-document reading, AA-LCR (Artificial Analysis 2026). AA-LCR asks 100 questions about long documents of 10,000 to 100,000 tokens, such as academic papers, legal documents, and company reports. To answer, a model must find information spread across the document and combine it, which is the same skill needed to code a syllabus, where the weight of a component and the way it is completed often sit on different pages.

On AA-LCR, at the medium reasoning level used for coding the syllabi, Gemini 3.7 Flash scores 83 percent. The most capable models from the two main frontier labs that were available when the syllabi were coded in August and September 2026, Claude Opus 5 from Anthropic and GPT-5.6 Sol from OpenAI, score 82 and 80 percent at the same level, and GLM-5.3 scores 80 percent at its highest reasoning level.[10] The models used in this paper therefore read long documents as well as the best models available at the time, at a small part of their cost.

3.2.2 Validation

To check the coding, a research assistant coded 500 syllabi by hand, without the help of generative AI: 250 from 2019 and 250 from 2022. The syllabi were drawn at random from all syllabi of each year, without any restriction. For each graded component, the research assistant recorded its weight and whether it is proctored, unproctored, or unclear. The same syllabi were then coded by Gemini 3.7 Flash, and the unproctored share was computed in the same way for both. The research assistant did not code task types, so this benchmark checks the unproctored share only.

Table 7 compares the codings, and across the 477 syllabi that all three coders could score, the correlation between the unproctored share from Gemini and from the research assistant is 0.82. The mean absolute difference is 0.079, the two codings are identical for 65 percent of syllabi, and they are within 10 percentage points of each other for 78 percent. Agreement is higher in 2019 than in 2022 (a correlation of 0.86 against 0.77), and much of the gap in 2022 comes from hybrid courses in the first year after COVID, which the research assistant often coded as fully in person. The syllabi with the largest disagreement, those where the two unproctored shares differ

by more than 0.20, were then reviewed by the author, and in 26 of the 31 syllabi that could be scored, the coding by Gemini was the correct one.

The second model, GLM-5.3. GLM-5.3 agrees with the research assistant almost as well as Gemini does, with a mean absolute difference of 0.083 and a correlation of 0.81 (Table 7), and the correlation between the two models on the benchmark is 0.93. On the 3,984 course sections coded by both models in the panels, the correlation between the two is 0.97 for both the unproctored share and AI exposure, and the difference between them does not change from year to year.

3.3 The AI Exposure Index

This section describes the three measures for the consistent courses of the 2018-2025 panel (Section 2.3). Each course is measured from its 2022 syllabi, the base year of the main specification, and, for comparison, from its 2019 syllabi. When a course has several sections in the base year, it receives the mean of its sections. Of the 1,027 consistent courses, 997 have at least one coded syllabus with grade weights in 2022, and 990 have one in 2019.

In the average course in 2022, 39 percent of the grade comes from AI-doable task types, 51 percent comes from unproctored work, and 30 percent comes from work that is both AI-doable and unproctored (Table 8). The measures vary widely across courses, and AI exposure is 1 percent at the 25th percentile and 50 percent at the 75th percentile, and about a quarter of courses (24 percent) have no exposure at all. Across the fall tracks used for estimation, the interquartile range of AI exposure is 45 percentage points (Table 10), and a move across it is the change in exposure used to describe the size of the effects.

2022 compared with 2019. Figure 2 compares the distribution of each measure in the two base years, and AI capability is almost the same in both years (a mean of 0.38 in 2019 and 0.39 in 2022), and the ranking of courses on this measure is stable, with a correlation of 0.77 between the two years for the 980 courses measured in both. The unproctored share, in contrast, rose from 0.37 to 0.51 between 2019 and 2022, as more graded work moved online during and after COVID, and its correlation across the two years is lower (0.60), so AI exposure also rose, from 0.23 to 0.30. The 2022 syllabi therefore describe courses as they were graded in the last year

before ChatGPT, after the changes made during COVID, while the 2019 syllabi describe them before COVID, which is why the 2019 measure is used as a check.

Two different dimensions. Figure 3 plots AI capability against the unproctored share for each course in 2022, and the two dimensions are only weakly related, with a correlation of 0.37. Many courses grade mostly unproctored work that AI cannot do, such as studio projects, lab work, and problem sets, and appear in the upper left of the figure. Other courses grade mostly AI-doable work in a proctored setting, such as essay exams or multiple-choice exams written in class, and appear in the lower right. AI exposure is high only for courses in the upper right, where the same components are both AI-doable and unproctored. For this reason, AI exposure is not simply a new name for either dimension: its correlation is 0.81 with AI capability and 0.64 with the unproctored share.

Differences across courses. Table 9 reports the mean of each measure by course level, class size, and field, and the largest differences are between fields. Courses in the humanities and arts and in the social sciences have a mean exposure of 0.37 and 0.38, against 0.18 in the natural sciences and mathematics and 0.15 in engineering. Most of this gap comes from AI capability, which is about twice as high in the humanities and social sciences as in the sciences and engineering (0.46 to 0.55 against 0.19 to 0.24), while the unproctored share differs much less across fields (0.42 to 0.57). Differences by course level are smaller, with a mean exposure of 0.31 in lower-division courses and 0.27 in upper-division courses. Courses with 100 or more students per section have a higher mean exposure (0.38) than courses with 25 to 99 students (0.24), because both their AI capability and their unproctored share are higher.

Departments. Figures 4 and 5 show the ten most exposed and the ten least exposed departments among the 34 departments with at least 10 measured courses. The most exposed departments are History (0.66), English (0.59), Social Work (0.55), and Philosophy (0.45), where most of the grade comes from written work done outside class. The least exposed are Mathematics (0.04), Studio Art (0.05), Chemical Engineering (0.08), and Physics (0.09). The two lowest departments show why both dimensions are needed: in Studio Art, 69 percent of the grade comes from unproctored work, but almost none of it is AI-doable, since the graded work is made by hand. In Mathematics, 40 percent of the grade is unproctored, but most of it consists of problem sets, which the free AI model of 2024-25 could not yet solve at a passing level. A measure based on

the unproctored share alone would place Studio Art well above the mean of all courses (0.51), and Mathematics about 11 points below it.

4 Empirical Strategy

This section describes how the effect of AI exposure on grades is estimated, using a design that compares the change in grades before and after the release of ChatGPT in November 2022 across courses with different levels of AI exposure. Exposure is measured from syllabi written before the release, so it describes how each course was graded when generative AI was not yet available to students. Section 4.1 presents the difference-in-differences model, and Section 4.2 presents the event study that follows from the same model.

4.1 Difference-in-Differences with Continuous Exposure

The estimation uses the panel of consistent courses for 2018 to 2025 (Section 2.3), with both fall and spring semesters. A course that is offered in every fall forms a fall track and a course that is offered in every spring forms a spring track, so that a course that qualifies in both semesters contributes one track for each semester. The unit of observation is a course in a semester, and time is measured by the semester, which combines the year and the term. Because a course is often taught in several sections and by different instructors in the same semester, all sections of a course in a semester are combined into one observation: the letter grades of all sections are pooled to build the share of A, the share of A- and mean grade points, and the standard deviation of grade points is computed within each section and then averaged across sections with weights equal to the number of letter grades (Section 2.2). The panel covers 16 semesters, from spring 2018 to fall 2025, and the estimation sample is a balanced panel of 909 tracks (493 fall tracks and 416 spring tracks) of 634 courses that have a coded 2022 syllabus and letter grades in all eight years, which gives 7,272 track-semester observations. The main estimates use the 493 fall tracks, with track and year fixed effects, and the estimates that add the 416 spring tracks, with track and semester fixed effects, are reported alongside them.

Treatment. The treatment is the AI exposure of each track, E_c , defined in equation (1) of Section 3. The capability of each graded component is binary, and a component counts as unproctored only when the syllabus shows that it is done outside the view of the instructor or a proctor; components whose format is unclear are counted as proctored (Section 3.1). Exposure is measured from the 2022 syllabi of the same term: fall 2022 syllabi for fall tracks and spring 2022

syllabi for spring tracks. When a course has several sections in 2022, its exposure is the mean of the exposure of its sections, weighted by the number of students in each section, as in Table 8, although the results for AI exposure are not sensitive to this weighting (Online Appendix Table B2, row 2d). This value is then held fixed in every semester, so that the treatment cannot respond to the release of ChatGPT and the changes that instructors made to their syllabi after 2022, which are studied in Section 6, do not enter it.

Model. The main specification relates the grade outcomes of each course to its exposure before and after the release of ChatGPT:

$$Y_{ct} = \alpha_c + \lambda_t + \beta \times E_c \times \text{Post}_t + \varepsilon_{ct} \quad (3)$$

Here Y_{ct} is a grade outcome of track c in semester t : the share of A, A- or B+, the cumulative shares (A- or above, B+ or above, B or above), mean grade points, the mean within-section SD, or the share of F or of D or F, built from all sections of the course as described above. The term α_c is a track fixed effect, so that the fall and the spring track of a course each have their own fixed effect, which absorbs all differences between courses that do not change over time, such as the field, the course level and the usual grading standard of the course, and it allows the fall and the spring versions of the same course to differ. The term λ_t is a semester fixed effect, with one effect for each term of each year, such as spring 2020 or fall 2024, which absorbs changes that affect all courses in the same semester, such as university-wide grading rules and the general rise in grades over time. The indicator Post_t equals 1 for every semester from spring 2023 on and 0 before. Because ChatGPT was released near the end of the fall 2022 term, fall 2022 is treated as the last semester before the release.[11] The separate effects of E_c and Post_t are absorbed by the two sets of fixed effects.

Interpretation. The coefficient β measures how much more the outcome changed after the release of ChatGPT in a track with full exposure ($E_c = 1$) than in a track with no exposure ($E_c = 0$). Few tracks are close to full exposure, so the text reports the effect of a realistic change in exposure: β multiplied by the interquartile range of exposure across tracks. Under the linear specification, this number is the predicted difference in the change of the outcome between a track at the 75th percentile of exposure and a track at the 25th percentile. Exposure is fixed in 2022 so that later changes in course design do not enter the treatment, and the coefficient β measures the differential change in grades after 2022 that is associated with exposure before the

release. Under the identifying assumptions below, this change can be attributed to the arrival of generative AI, and it may include both the use of AI by students and any change that instructors made in response. An effect that is common to all courses is absorbed by the semester fixed effects, so it is not part of the estimate. Standard errors are clustered at the course level, so the fall and the spring track of the same course form one cluster, and each track-semester observation has equal weight.

Identification. The main assumption is parallel trends: without ChatGPT, grades in courses with high and low exposure would have followed the same path over time. With a continuous treatment, Callaway, Goodman-Bacon and Sant'Anna (2024) show that, under this assumption alone, the coefficient in equation (3) measures differences in the change of grades across levels of exposure, which can mix the effect of exposure with differences in how courses at each level respond. The assumption also requires that other changes after 2022, such as the recovery from COVID or changes in the students who take each course, did not affect grades differently in courses with high and low exposure. To read β as the effect of a higher level of exposure, a stronger version of the assumption is needed: courses with different levels of exposure would have responded in the same way to the same level of exposure. The parallel trends assumption cannot be tested directly, but the event study in Section 4.2 shows whether grades in high- and low-exposure courses moved together before 2023.

Other measures of the treatment. Equation (3) is also estimated with AI capability in place of E_c , and the two measures differ in the graded work that they count. AI capability counts all graded work that AI can do, including work done under a proctor, where AI can change grades mainly through the learning effect, although student effort and instructor responses can also play a role. AI exposure keeps only the part of this work that students complete without a proctor, where the substitution effect is also possible because AI can produce the graded work itself. If the unproctored share matters for grades, the two measures should move different parts of the grade distribution, so Section 5 reports the estimates for AI capability and for AI exposure side by side. This comparison describes how the estimates differ across kinds of graded work, but it does not separate the learning effect from the substitution effect, because the two measures overlap. As a robustness check, equation (3) is also estimated with exposure measured from the 2019 syllabi of the same term, that is, fall 2019 syllabi for fall tracks and spring 2019 syllabi for

spring tracks, built with the same rules as the 2022 measure. The 2019 syllabi describe how courses were graded before the changes made during COVID, when the unproctored share was lower (Section 3.3), and they were written three years before the release of ChatGPT. The sample is limited to courses with a coded 2019 syllabus, while the outcomes, the fixed effects and the clustering stay the same. If the two base years give similar estimates, the results do not depend on the year in which exposure is measured. Because the 2019 syllabi were written before the COVID shock to grading and the 2022 syllabi after it, similar estimates make it less likely that the results reflect the changes in course design made during COVID. The estimates with the 2019 base year are summarized in Section 7 and reported in Online Appendix B, together with the other robustness checks.

4.2 Event Study

Equation (3) gives one average effect for the whole period after the release, so, to show how the effect changes over time, and to check the parallel trends assumption, the single term $Post_t$ is replaced with one indicator for each year:

$$Y_{ct} = \alpha_c + \lambda_t + \sum_{k \neq 0} \beta_k \times E_c \times D_{kt} + \varepsilon_{ct} \quad (4)$$

The indicator D_{kt} equals 1 if semester t falls in year $2022 + k$, so k counts years from the base year 2022 and runs from -4 (2018) to 3 (2025). The outcomes, the fixed effects, the treatment, the sample and the clustering are the same as in equation (3).

Reference period. The year $k = 0$ is left out, so each β_k is measured relative to 2022. For a fall track the reference semester is fall 2022, and for a spring track it is spring 2022. These are the semesters in which exposure is measured. Spring 2022 ended before the release of ChatGPT, while fall 2022 was close to its end when ChatGPT was released on November 30, 2022, so fall 2022 is treated as a pre-release semester on the assumption that ChatGPT had little effect on its final grades. Event time is counted in years rather than in semesters because a spring track is never observed in a fall semester, and a fall track is never observed in a spring semester. The specification therefore uses 2022 as the reference year, comparing each track with its own term in 2022, and the fall and the spring semester of the same year share one coefficient. This pooling is a modeling choice rather than a necessity, and the event studies reported in Section 5 use the fall tracks alone, so they do not impose it. With this timing, $k = 1$ is the first semester after the release for both tracks, spring 2023 for spring tracks and fall 2023 for fall tracks.

Pre-trends. The coefficients β_{-4} to β_{-1} measure how grades in high- and low-exposure courses moved apart before the release of ChatGPT, and under parallel trends they should be close to zero. Estimates close to zero, however, do not prove that the trends were parallel, because such comparisons can have low power (Roth 2022).

Effects after the release. The coefficients β_1 to β_3 measure the effect in 2023, 2024 and 2025. They show whether the effect appears at once or grows over time, for example as more students adopt AI tools and as the tools improve. Section 5 shows these estimates in event-study figures with pointwise 95 percent confidence intervals and the reference year marked.

5 Results

This section shows how grades changed after the release of ChatGPT in courses where more of the grade comes from work that AI can do. The main result is that grades rose in these courses, and the rise was largest at the top of the distribution, where more students received an A and fewer received an A- or a B+.

Section 5.1 reports the results for AI capability, and Section 5.2 shows how they change when only the work that students complete without a proctor is counted, as in AI exposure. Section 5.3 shows which letter grades became less common, and Section 5.4 shows how grades changed in each year before and after the release. Section 7 shows that the main results hold under other ways of measuring the treatments and of choosing the sample. To make the size of each effect easy to read, the text compares a course at the 75th percentile of each measure with a course at the 25th percentile.

5.1 AI Capability and Grades

Columns (1) and (2) of Table 10 show that grades rose more after the release of ChatGPT in courses where a larger share of the grade comes from task types that AI can do. In column (1), the share of A grades rose by 0.085 for a one-unit change in AI capability. A move from the 25th to the 75th percentile of AI capability therefore raised the share of A by 5.6 percentage points, which is about 13 percent of the mean share of A before 2023 (44.6 percent). The higher share of A came together with lower shares of the grades just below it. For the same move in AI capability, the share of A- fell by 2.2 percentage points and the share of B+ fell by 1.1 percentage points.

Because the shares of A- and B+ fell by less than the share of A rose, more students received high grades overall. For a move from the 25th to the 75th percentile of AI capability, the share of students with A- or above rose by 3.4 percentage points (Table 10, Panel B). For the same move, the share with B+ or above rose by 2.3 percentage points, and the share with B or above rose by 1.3 percentage points. Mean grade points rose by 0.067 grade points for a one-unit change in AI capability, which equals 0.04 grade points for a move from the 25th to the 75th percentile.

Grades in courses with higher AI capability also became somewhat more compressed, because the mean within-section SD fell by 0.036 grade points for a one-unit change. This estimate is significant at the 10 percent level and equals 0.02 grade points for a move from the 25th to the 75th percentile of AI capability. The shares of F and of D or F did not change (Panel D), so the rise in grades did not come from fewer students failing. It was matched by falls in the shares of A- and of grades in the B range.

When the spring tracks are added, in column (2) of Table 10, the estimates for the share of A, the cumulative shares and mean grade points are similar in size. In this column, the mean within-section SD fell by 0.049 grade points for a one-unit change, an estimate that is significant at the 1 percent level.

5.2 AI Exposure and Grades

AI exposure counts only the part of the AI-doable work that students complete without a proctor, so it adds the proctoring condition to AI capability. Columns (3) and (4) of Table 10 report the estimates for this measure, using the same samples as columns (1) and (2). In column (3), the share of A rose by 0.100 for a one-unit change in AI exposure, which equals 4.5 percentage points for a move from the 25th to the 75th percentile. For the same move, the share of A- fell by 1.8 percentage points and the share of B+ fell by 1.1 percentage points. These changes at the top of the distribution are also found with the 2019 syllabi and in the longer panel for 2015 to 2025 (Online Appendix Table B2, rows 1 and 5).

For the rest of the grade distribution, the evidence is much weaker for AI exposure than for AI capability. In column (3), the share of students with A- or above rose by 2.7 percentage points and mean grade points rose by 0.035 grade points for a move from the 25th to the 75th percentile. Over the same move, the mean within-section SD fell only a little, and this change is

not statistically significant. With the 2019 syllabi, the estimates for the cumulative shares, mean grade points and the mean within-section SD are all close to zero and are not significant, and mean grade points rose by only 0.001 grade points for a move from the 25th to the 75th percentile of AI exposure (Online Appendix Table B2, row 1). In the 2015-2025 panel, the same estimate is 0.029 grade points and is significant only at the 10 percent level (Online Appendix Table B2, row 5). The event study in Figure 8 shows that the estimate for mean grade points in column (3) of Table 10 comes mainly from lower grades in high-exposure courses before 2022, not from a rise after the release.

The two measures therefore show different patterns, and the difference comes from adding the proctoring condition to AI capability. With AI capability, grades rose across the upper half of the distribution, mean grade points rose, and the mean within-section SD fell. With AI exposure, the clear and stable result is a shift from A- and B+ grades to A grades, while the effects on mean grade points and on the mean within-section SD are no longer clear. The rise in mean grade points is thus linked to task types that AI can do, whether the work is proctored or not. The part of this work that students complete without a proctor mainly moves grades from just below the top to the top. As with AI capability, the shares of F and of D or F did not change in columns (3) and (4) of Table 10.

5.3 Changes across the Grade Distribution

Figure 6 shows the effect of AI exposure on the share of each letter grade, with one regression of equation (3) for each grade, using the fall tracks and the 2022 syllabi. Because the shares of all letter grades add up to one, the estimates show which grades became less common as the share of A rose. For a one-unit change in AI exposure, the share of A rose by 10.0 percentage points. In contrast, the share of A- fell by 4.0 percentage points, the share of B+ by 2.5 percentage points, and the share of B- by 1.7 percentage points. The share of B fell by 1.3 percentage points, but this estimate is not statistically significant. The shares of C, D and F grades did not change, and no estimate for these grades is larger than 0.5 percentage points in absolute value.

The rise in the share of A was therefore matched almost entirely by falls in the shares of A- and of grades in the B range. This is the pattern of grade compression at the top of the distribution that is described in the introduction. Because A is the highest grade that an undergraduate can

receive at UT Austin, more students now reach the ceiling of the grade scale. As a result, grades separate fewer groups of students in the upper part of the distribution.

5.4 Event Studies

Figures 7 and 8 show the event studies for AI capability and for AI exposure, estimated with equation (4) for the fall tracks and the 2022 syllabi. Each point compares the outcome in a given year with the outcome in 2022, which is the reference year. The points before 2023 show whether courses with high and low values of the measure followed the same path before the release of ChatGPT, and the points from 2023 on show how the effect changed over time.

For AI capability, the share of A followed the same path in courses with high and low AI capability before 2023 (Figure 7, Panel A). The estimates for these years lie between -0.031 and 0.009 and are not statistically different from zero. After the release of ChatGPT, the share of A rose by 0.072 in 2023, by 0.084 in 2024 and by 0.091 in 2025 for a one-unit change in AI capability. The effect therefore appeared in the first year after the release and grew slowly over the next two years. In the same years, the share of A- fell by 0.033 to 0.051 , and the share of students with A- or above rose from 0.021 in 2023 to 0.059 in 2025. Before 2023, the estimates for these two outcomes are small and are not significant at the 5 percent level. The share of B+ moved less steadily before 2023, so its fall after 2022 should be read with more caution than the changes in the shares of A and A-.

Mean grade points and the mean within-section SD follow the same pattern for AI capability (Figure 7, Panel C), and before 2023 the estimates for both outcomes are small and are not statistically significant. After the release, mean grade points rose in each year and reached 0.073 grade points in 2025, while the mean within-section SD fell and reached -0.045 grade points in 2025. These yearly estimates are less precise than the single estimate in Table 10, because each of them uses the grades of only one year after the release. The shares of F and of D or F show no change after 2022 (Panel D).

For AI exposure, the share of A also rose after the release of ChatGPT, by 0.068 in 2023, by 0.056 in 2024 and by 0.065 in 2025 (Figure 8, Panel A). These estimates are for a one-unit change in AI exposure, and the share of A- fell in each of the same years. Before 2023, the estimates for the share of A are negative, between -0.066 and -0.037 , but only the estimate for

2019 is significant at the 5 percent level. The rise in the share of A after the release is clear for AI exposure, although part of the difference-in-differences estimate reflects the negative estimates before 2023, which disappear when exposure is measured from the 2019 syllabi.

The other outcomes do not show a clear change after the release for AI exposure. For mean grade points, the estimates before 2023 range from -0.108 to -0.046 grade points, while the estimates from 2023 on range from -0.023 to 0.038 and are close to zero (Figure 8, Panel C). The positive estimate for mean grade points in column (3) of Table 10 therefore comes from lower grades in high-exposure courses before 2022, not from a rise after the release. The cumulative shares in Panel B follow the same pattern, with negative estimates before 2023 and estimates close to zero after the release. One possible reason is that the 2022 syllabi were written after the move to online and take-home work during COVID (Section 3.3). Courses with high exposure in 2022 may therefore include courses whose grading changed, and whose grades rose, between 2020 and 2022. This reading agrees with the estimates with the 2019 syllabi, which show no change in mean grade points after the release (Online Appendix Table B2, row 1).

The event studies therefore support the results in Table 10, and they show the same difference between the two measures. For AI capability, the shares of A and of high grades rose after 2022, mean grade points rose, and the mean within-section SD fell, with no clear difference between courses before the release. For AI exposure, the clear change after 2022 is limited to more A grades and fewer A- grades at the top of the distribution.

6 Instructor Response and Mechanisms

After the arrival of GenAI tools in November 2022, grades rose more in courses where a larger share of the grade comes from work that AI can do. The most robust change was a shift from A- and B+ grades to A grades, while mean grade points rose by only 0.04 grade points for a move from the 25th to the 75th percentile of AI capability (Table 10). When only the work that students complete without a proctor is counted, as in AI exposure, the rise in the share of A was not larger, at 4.5 against 5.6 percentage points for the same move (Table 10), although per one-unit change the estimate is larger for AI exposure (0.100 against 0.085). With AI exposure, the effects on mean grade points and on the mean within-section SD were also no longer clear. If students mainly used AI to complete work done without a proctor, the effects should be larger for AI exposure than for AI capability, so this pattern needs an explanation.

One explanation is the change in grading during COVID, and a second possible explanation is that instructors responded to GenAI by changing their courses. Instructors may have added rules on AI use, moved grade weight away from work that AI can do, or proctored more of this work, and each of these changes limits how much AI can raise grades. Because the main estimates hold AI exposure fixed at its 2022 level, such changes after 2022 are already part of them and would make the estimated effects smaller than the effects with no response. The analysis below asks whether instructors changed their syllabi after the arrival of GenAI, using the panel of consistent instructors (Table 5) rather than the panel of consistent courses that is used for the grade results. An instructor is in this panel if the instructor has at least one syllabus in every fall or in every spring from 2018 to 2025, which gives 529 instructors and 20,585 sections, whose syllabi were coded with GLM-5.3 under the same rules as the main measure.

Instructors could respond to GenAI in several ways, but only the responses that are written in the syllabus can be measured with these data. Three responses can be measured: the rules on AI use stated in the syllabus, the share of the grade that comes from each type of work, and whether graded work is done with a proctor. Other responses leave no trace in the syllabus, because instructors may make the same assignments harder for AI, use AI detectors to enforce their rules, or stop curving scores upward at the end of the semester. Some instructors curve scores upward at the end of the semester, for example by raising the highest score to 100 and scaling the other scores in the same way, and they may stop doing so when scores are already high after the arrival of GenAI. The results in this section say nothing about these responses, so they show only part of what instructors may have done. The measured responses are studied in order of their cost to the instructor, from the cheapest to the most costly. Section 6.1 describes the rules on AI use and whether the instructors who adopted them also changed the design of their courses, Section 6.2 describes the grade weights, and Section 6.3 describes proctoring.

6.1 Rules on AI Use

The cheapest response to GenAI is a rule in the syllabus that states whether and how students may use AI, which is called an AI policy in what follows. AI policies are measured from the text of each syllabus, which GLM-5.3 read for any mention of artificial intelligence, GenAI tools such as chatbots, or large language models. Each syllabus received one of five codes: a syllabus has a prohibited policy, or ban, when AI tools are not allowed for course work. It has a

conditional policy when AI use is allowed only under stated conditions, such as use for brainstorming but not for writing, use with a citation, or use in some assignments only. A syllabus has a permissive policy when AI tools are allowed or built into the course without real limits, an unclear policy when AI is mentioned but the rule is vague, and no policy when AI is not mentioned. Any code other than no policy counts as an AI policy, and each syllabus for one course in one term is counted once, even when several sections share it, which gives 11,909 syllabi from 2018 to 2025. Before 2023, between 1.5 and 2.4 percent of syllabi in each year are coded as having a policy, although none of these syllabi contains a word related to AI. These codes come mostly from bans on online translators in language courses and from general rules on academic honesty that the model read as bans on AI, so this share gives the level of error in the measure.

AI policies spread quickly after the arrival of GenAI, but by 2025 about half of all syllabi still had none. The share of syllabi with an AI policy rose from 2.4 percent in 2022 to 21.1 percent in 2023, 41.2 percent in 2024 and 50.4 percent in 2025 (Figure 9, Panel A). The share for 2023 is the lowest of the three years partly because it includes the spring semester, whose syllabi were written only a few weeks after the release, when few instructors had added a policy. In 2023, bans were the most common type, at 12 percent of all syllabi against 7 percent for conditional policies, and from 2023 to 2024 the shares of both types rose by about 10 percentage points. From 2024 to 2025, the share of syllabi with a ban stayed at about 21 percent, so almost all of the growth came from conditional policies, whose share rose from 17 to 26 percent of all syllabi (Figure 9, Panel A). As a result, the mix of policies moved away from bans, which made up 57 percent of all policies in 2023 and 42 percent in 2025, while conditional policies rose from 35 to 51 percent (Figure 9, Panel B). Permissive policies made up at most 3.5 percent of all policies in any year, so instructors who wrote a policy almost always chose to limit the use of AI rather than to encourage it.

AI policies were also more common in courses where AI could do a larger part of the graded work before the arrival of GenAI. To show this, each pair of instructor and course is placed in one of three groups by its AI capability in 2022, the share of the course grade that comes from the task types that AI can complete. These task types are written prose, objective items such as multiple-choice questions, and code or data. In the 2022 syllabi, a course has low AI capability

when less than one third of its grade comes from these task types, middle when one third to two thirds does, and high when more than two thirds does. The groups are fixed at 2022, so that they do not change when instructors change their syllabi after the release, and the 1,210 pairs of instructor and course taught in 2022 give 8,657 of the 11,909 syllabi. Before 2023, the share of syllabi with a policy was below 5 percent in all three groups, in line with the error in the coding (Figure 10). After the release, this share rose in every group, but it rose faster in the middle and high groups. In 2025, 59 percent of syllabi in the high group and 53 percent in the middle group had an AI policy, against 42 percent in the low group (Figure 10). The gap between the high and low groups grew from 12 percentage points in 2023 to 17 percentage points in 2024 and 2025, and it is significant at the 1 percent level in each year, with standard errors clustered by instructor. Instructors were therefore more likely to adopt a policy when AI could do more of the graded work in their course, although the shares in the middle and high groups are not statistically different from each other.

An AI policy changes only the rules of a course, so a natural question is whether the instructors who adopted the strictest policy, a ban, also changed the design of their courses. Figure 11 compares the change in four parts of the syllabus from 2022 to 2025 between the courses with a ban in their 2025 syllabus and all other courses. It uses the 742 pairs of instructor and course that were taught in both years, of which 159 have a ban in 2025, while the other 583 have a conditional, permissive or unclear policy, or no policy at all. Both groups moved grade weight away from unproctored work and toward proctored exams, but the move was larger in the courses with a ban. In these courses, the unproctored share of the grade fell by 10.6 percentage points, against 5.1 percentage points in the other courses (Figure 11). This gap of 5.5 percentage points is statistically significant at the 5 percent level ($p = 0.04$), with standard errors clustered by instructor, even though the confidence intervals of the two groups overlap slightly. For AI exposure, AI capability and the share of proctored exams, the gaps between the two groups are not statistically significant at the 10 percent level. The larger fall in the unproctored share should, however, be read together with the years before 2022, because the courses with a ban had also added more unproctored work during COVID. When the change is measured from 2019 to 2025 instead of from 2022 to 2025, the unproctored share rose by 8.8 percentage points in these courses and by 7.7 percentage points in the other courses, and the gap of 1.1 percentage points is not significant. The larger fall after 2022 therefore mostly reversed a larger rise during COVID,

so a ban did not come with a course design that relies less on unproctored work than before the pandemic.

Taken together, the syllabi show that instructors responded to GenAI mainly through rules, since half of all syllabi had an AI policy by 2025 and policies were more common where AI could do more of the graded work. These policies changed what students were told rather than how their work was graded, because even the courses with a ban did not reduce the unproctored part of the grade more than other courses once the COVID years are taken into account. Whether a policy limits the use of AI depends on how it is enforced, which the syllabus does not show, so AI policies alone may have done little to limit the effect of AI on grades. The rest of this section asks whether instructors changed the grade weights and the proctoring of their courses more directly.

6.2 Grade Weights

A more costly response than a written policy is to change what the grade is made of, so that less of it comes from the task types that AI can do. To measure this response, the share of the grade that comes from each task type is computed for every syllabus. The shares are then compared within the same pair of instructor and course, so that changes in the courses that an instructor teaches do not drive the results. The analysis uses 9,604 observations of 2,784 pairs of instructor and course from 2018 to 2025, where each observation is the mean over the syllabi of a pair in a year. The change in each share is estimated with an event study, a regression of the share on a fixed effect for each pair and an indicator for each year except 2022, with standard errors clustered by instructor. Each year coefficient is then the average change in the share from 2022 to that year within the same pair, and the coefficients for the years before 2023 show whether the shares were already changing before the release. Because every course faced GenAI at the same time, there is no comparison group, so the estimates describe the average change after the release rather than a causal effect of AI. Unlike the unproctored share, AI capability barely moved during COVID, from 43 percent of the grade in 2019 to 44 percent in 2022, so the comparison with 2022 is not distorted by the move to online teaching.

Grade weight moved slowly away from the task types that AI can do after the arrival of GenAI, and the move grew a little each year. AI capability did not change in the years before 2023, and it then fell by 0.6 percentage points in 2023, 1.3 percentage points in 2024 and 1.7 percentage

points in 2025, relative to 2022 (Figure 12, Panel A). Only the fall in 2025 is significant at the 5 percent level, and it equals about 4 percent of the mean AI capability in 2022. At the same time, the share of work that is hard for AI, such as participation, oral work and labs, rose by 1.2 percentage points in 2023 and 2024 and by 1.8 percentage points in 2025 (Figure 12, Panel B). Each of these three estimates is significant at the 5 percent level, while the estimates for the years before 2023 are slightly negative but not significant, so part of the rise may continue a small trend that started before the release.

Most of the change came from one task type on each side, because from 2022 to 2025 the share of written prose fell by 1.8 percentage points and the share of participation rose by 1.0 percentage point (Figure 13). The shares of objective items and of code or data did not change, while oral work and studio or lab work rose by about half a percentage point each, although these two estimates are not statistically significant. The shares of quantitative problem sets and of components whose task is not stated did not change either. Instructors therefore replaced some essays and papers with participation, while they kept the quizzes, tests and coding work that make up the rest of the work that AI can do.

These changes are too small to change the reading of the main estimates, which hold AI capability fixed at its 2022 level. The fall of 1.7 percentage points in AI capability is about 3 percent of its interquartile range in the panel of consistent courses, which is 66 percentage points (Table 10). It is therefore small next to the differences between courses that the main estimates rely on, and the 2022 measure still describes well the work that students were graded on after the release. Instructors did change the grade weights of their courses after the arrival of GenAI, but the change was slow and small.

6.3 Proctoring

The most costly response is to move graded work into a proctored setting, for example by replacing a take-home essay with an in-class exam, so that AI cannot be used to complete it. The unproctored share of the grade and AI exposure are measured with the same rules as the main measure, so that components with an unclear format and work done live online count as proctored. Both measures are compared within the same pair of instructor and course, as in the analysis of grade weights. Proctoring, however, was also changed by COVID, because the move to online teaching in 2020 and 2021 turned much proctored work into take-home work, and in

2022 the unproctored share was still about 14 percentage points above its 2019 level. A fall in the unproctored share after 2022 could therefore be a response to GenAI or simply a return to the practice before COVID. For this reason, the event studies in Figure 14 use 2019, the last year before COVID, as the reference year.

To separate the two explanations, the pairs of instructor and course are split by what happened to their proctoring during COVID. The first group contains the 1,388 pairs whose syllabi report in-person teaching in every year in which they were taught, while the second contains the 278 pairs whose unproctored share in 2022 was within 5 percentage points of their 2019 level. The third group contains the 347 pairs whose unproctored share in 2022 was still more than 5 percentage points above their 2019 level. In this third group, the unproctored share was 33 percentage points above its 2019 level in 2022, and it then fell to 24, 22 and 18 percentage points above 2019 in 2023, 2024 and 2025 (Figure 14, Panel A). The other two groups show no change after 2022, because the courses that had already returned to their 2019 level stayed there. The courses that were always taught in person kept an unproctored share between 5 and 7 percentage points above 2019 in every year from 2022 to 2025. AI exposure follows the same pattern, with a fall after 2022 only in the third group, from 16 to 8 percentage points above its 2019 level (Figure 14, Panel B). The fall in unproctored work after 2022 therefore took place only in courses that still had a gap to close after COVID, and no group of courses reduced its unproctored work below the level of 2019.

A response to GenAI should also be larger where more of the grade could be earned with AI and without a proctor before COVID. Figure 15 tests this with an event study in which each year indicator is multiplied by the 2019 value of the outcome, called the dose. Each coefficient then shows how much more the outcome changed relative to 2022 in a course with a higher 2019 value. Only the years from 2020 on are used, because the years 2018 and 2019 would compare the dose with itself and create a mechanical relation. The estimates are shown both for the same pair of instructor and course and for the same instructor in any course, with fixed effects for each year and for each pair or instructor, and standard errors clustered by instructor. None of the yearly coefficients after 2022 is statistically different from zero, and the difference-in-differences estimates for the years from 2023 on are small and not significant in either comparison (Figure 15). For a move from the 25th to the 75th percentile of the dose, the estimates for the same pair

imply an extra change of 2.0 percentage points in the unproctored share after 2022 and a fall of 0.9 percentage points in AI exposure. The courses and instructors with the most work that AI could do without a proctor before COVID therefore did not move more of this work into proctored settings after the arrival of GenAI.

Together, these results suggest that proctoring did not respond to GenAI beyond the return from COVID, and that the fall in AI exposure after 2022 mostly reflects the end of the move to take-home work during COVID rather than a response to AI. This has one consequence for the main estimates, which measure AI exposure in the 2022 syllabi, because in some courses a high exposure in 2022 came from take-home work that was added during COVID and removed later. In these courses, the 2022 measure overstates exposure in the years after 2022 and adds noise to the estimates. The measure from the 2019 syllabi is not affected by this problem, and with it the share of A still rose by 2.1 percentage points for a move from the 25th to the 75th percentile of AI exposure (Online Appendix Table B2). Across this section, instructors responded to GenAI mainly through rules on AI use, changed the grade weights of their courses only a little, and did not change proctoring beyond the return from COVID. Their response is therefore too small to explain why the effect of AI on grades was modest.

7 Robustness Checks

The main estimates rest on several choices about how the treatment is measured and which courses enter the sample, and Appendix B of the Online Appendix repeats the estimation with each of these choices changed, one at a time. The checks measure the treatment from the 2019 syllabi, replace the binary coding of AI capability with benchmark scores, code quantitative problem sets as AI-doable, give each 2022 section of a course the same weight, count components with indeterminate proctoring as half unproctored, keep only in-person courses, and use the longer panel for 2015 to 2025. Each estimate is reported as the effect of a move from the 25th to the 75th percentile of the treatment in that sample (Online Appendix Tables B1 and B2), and the full regression tables are in Online Appendix C.

For AI exposure, the shift from A- and B+ grades to A grades appears in every check. In the fall tracks, an interquartile move in AI exposure raised the share of A by between 2.1 and 4.9 percentage points across the checks, against 4.5 percentage points in the main estimate, and it lowered the shares of A- and B+ in every check (Online Appendix Table B2). The smallest effect

comes from the 2019 syllabi, which describe courses before COVID, while coding quantitative problem sets as AI-doable lowers the effect to 3.5 percentage points and giving each section the same weight leaves it at 4.5 percentage points. The rise in mean grade points and the fall in the mean within-section SD are smaller and less precise in several checks, and they disappear when exposure is measured from the 2019 syllabi. The negative estimates before 2023 in the event study for AI exposure are specific to the 2022 measure in the full sample, because they are absent with the 2019 syllabi and in the in-person courses.

Taken together, the checks support the view that the main effect of generative AI on grades is a move from just below the top grade to the top grade, and that this effect is linked to graded work that AI can do and that students complete without a proctor.

8 Conclusion

This paper examines changes in undergraduate grades following the release of ChatGPT, using linked grade distributions and syllabi from the University of Texas at Austin. It develops a measure of AI exposure that combines the task content of graded work, its weight in the final grade, and whether students complete it without a proctor. Measuring exposure from syllabi written before ChatGPT's release allows the analysis to compare subsequent changes in grades across courses with different assessment structures.

The most robust finding is an increase in the concentration of grades at the top of the distribution. In the fall tracks, an interquartile difference in baseline AI exposure is associated with a 4.5-percentage-point larger increase in the share of A grades after 2022, about 10 percent of its pre-2023 mean (44.6 percent). This increase is accompanied by declines of 1.8 and 1.1 percentage points in the shares of A- and B+ grades, while there is little evidence of changes in the shares of low grades. The pattern at the top persists across alternative exposure measures and samples, although its magnitude is smaller when exposure is constructed from 2019 syllabi. Estimates for mean grade points and the mean within-section SD are less stable across specifications, so the evidence points most consistently to greater concentration at the grade ceiling.

Instructors' documented responses to generative AI consisted primarily of changes in the policies stated in their syllabi. By 2025, approximately half of the syllabi in the instructor panel contained

an AI policy, with policies more common in courses where AI could perform a larger share of the graded work. Assessment design changed more modestly, since within instructor-course pairs the grade weight assigned to task types classified as within AI's capabilities declined by only 1.7 percentage points between 2022 and 2025. Reductions in unproctored work were concentrated in courses that had not returned to their pre-pandemic assessment practices, a pattern consistent with continuing pandemic recovery. These findings describe responses visible in syllabi; they do not capture changes in enforcement, assignment difficulty, or grading practices that instructors did not document.

The findings raise questions about the information that academic grades convey to employers and graduate programs. Greater concentration at the upper bound may weaken GPA's signaling value if students with different levels of proficiency increasingly receive the same grades. Whether this concentration reduces the information that grades carry about proficiency or later productivity cannot be tested with section-level data. These records also cannot distinguish improved learning from AI assistance in producing assessed work, or identify which individual students moved between grade categories.

These limitations, together with the speed at which AI tools improve, suggest several directions for further research. The estimates describe the first generation of generative AI tools, which students used from 2023 to 2025, and the effects of later, more capable tools may be larger. Linking assignment-level outcomes to independent measures of proficiency would help determine whether gains on unproctored work carry over to performance without AI assistance. Student-level records would also permit analysis of cumulative GPA, academic progression, and subsequent labor-market outcomes. Extending the syllabus-based approach to other institutions and later cohorts would show how these relationships vary with grading practices, AI capabilities, and instructor adaptation. The central question is whether assessment can continue to distinguish students' knowledge and skills as the tools available to produce graded work change.

Footnotes

1. Calculations in this paper from the National Postsecondary Student Aid Study, 2000 and 2020.
2. Humanity's Last Exam has 2,500 closed-ended questions in more than 100 subjects, written by about 1,000 experts. Scores are from the AI Dashboard of the Center for AI Safety and Scale AI (lastexam.ai), accessed on September 30, 2026. GPT-4o, the model whose benchmark scores define AI capability in this paper, answered only 2.7 percent of these questions correctly, so the test measures expert-level ability far above what an undergraduate assignment requires. Figure D1 in the Online Appendix shows the score of each OpenAI model over time.
3. Texas Education Code, Section 51.974, added by H.B. 2504, 81st Texas Legislature (2009), enrolled version. The law also requires a departmental budget report, if available.
4. The dashboard does not show a course when student privacy cannot be protected because of its size. For undergraduate sections, this rule removes sections with fewer than 10 students (see below).
5. Calculations in this paper from the download log of the UT Austin course documents site. Counts cover fall and spring semesters. CV counts are section-linked records, not distinct instructors. Of the 131,026 sections, 11 have a CV but no syllabus. Subject areas are course prefixes.
6. Calculations in this paper. Sections in fall and spring semesters, 2015 to 2025, are matched by year, semester, and unique section number. Graduate and professional sections are those with the last two digits of the course number from 80 to 99, and all courses with the prefixes LAW, PHR, and MED (law, pharmacy, and medicine). Lower-division sections have last two digits from 01 to 19, and upper-division sections from 20 to 79, following the UT Austin course numbering system. More than half of the undergraduate grade sections without a syllabus are laboratory sections (course titles with "laboratory", "lab", "practice", or "experiment"), which are likely covered by the syllabus of the related lecture course.
7. In Scarfe et al. (2024), essays and short answers written by GPT-4 were submitted into the exam system of a UK university without the graders knowing. They received a higher grade than a randomly chosen student in 83 percent of comparisons, and 94 percent of

them were not detected. The study covers one field (psychology) and 33 submissions. GPT-4o, the free model in 2024-25, is a newer model of the GPT-4 family (OpenAI 2024b).

8. Free users who reached the usage limit of GPT-4o were moved to GPT-4o mini, which also passes the line for both types (82.0 percent on MMLU and 87.2 percent on HumanEval; OpenAI 2024a).
9. On MATH, a set of competition problems, GPT-4o scored 75.9 percent (Hendrycks et al. 2021b; OpenAI 2024a). MATH problems are shorter and more standard than university problem sets, so U-MATH is the benchmark closer to an undergraduate assignment. Problem sets reach the 70 percent line on the free tier only in 2025-26, when reasoning models became free.
10. Scores are from AA-LCR version 1.1, accessed on October 4, 2026, for Gemini 3.7 Flash (medium), Claude Opus 5 (medium), and GPT-5.6 Sol (medium); Claude Opus 5 and GPT-5.6 Sol were released in July 2026. For GLM-5.3, the benchmark reports only the highest reasoning level (max), while the coding runs used the high level. Newer frontier models, released on September 22, 2026, after the syllabi were coded, are not included.
11. ChatGPT was released on November 30, 2022 (OpenAI 2022). At that point the fall 2022 term was in its last weeks, and the syllabi used to measure exposure had been written before the term began.

References

- Artificial Analysis. 2026. "AA-LCR: Artificial Analysis Long Context Reasoning Benchmark, version 1.1." artificialanalysis.ai/evaluations/artificial-analysis-long-context-reasoning (accessed October 4, 2026).
- Bastani, Hamsa, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. 2025. "Generative AI without Guardrails Can Harm Learning: Evidence from High School Mathematics." *Proceedings of the National Academy of Sciences* 122 (26).
- Callaway, Brantly, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna. 2024. "Difference-in-Differences with a Continuous Treatment." NBER Working Paper 32117.
- Center for AI Safety and Scale AI. 2026. "Humanity's Last Exam." AI Dashboard, lastexam.ai. Accessed September 30, 2026.
- Chan, William, Li Hao, and Wing Suen. 2007. "A Signaling Theory of Grade Inflation." *International Economic Review* 48 (3): 1065-1090.
- Chen, Mark, et al. 2021. "Evaluating Large Language Models Trained on Code." arXiv 2107.03374.
- Chernyshev, Konstantin, et al. 2024. "U-MATH: A University-Level Benchmark for Evaluating Mathematical Skills in LLMs." arXiv 2412.03205.
- Chirikov, Igor. 2026. "Artificial Intelligence and Grade Inflation." CSHE Higher Education Working Paper Series 26-3, University of California, Berkeley.
- Contractor, Zara, and Germán Reyes. 2026. "Generative AI in Higher Education: Evidence from an Elite College." arXiv 2508.00717.
- Denning, Jeffrey T., Eric R. Eide, Kevin J. Mumford, Richard W. Patterson, and Merrill Warnick. 2022. "Why Have College Completion Rates Increased?" *American Economic Journal: Applied Economics* 14 (3): 1-29.
- Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2024. "GPTs Are GPTs: Labor Market Impact Potential of LLMs." *Science* 384 (6702): 1306-1308.
- Fair, Ray C. 2023. "Grade Report Update: 2022-2023." Yale University, November.
- Felten, Edward, Manav Raj, and Robert Seamans. 2021. "Occupational, Industry, and Geographic Exposure to Artificial Intelligence: A Novel Dataset and Its Potential Uses." *Strategic Management Journal* 42 (12): 2195-2217.
- Geerling, Wayne, G. Dirk Mateer, Jadrian Wooten, and Nikhil Damodaran. 2023. "ChatGPT Has Aced the Test of Understanding in College Economics: Now What?" *The American Economist*. <https://doi.org/10.1177/05694345231169654>
- Hansen, Anne Toft, Ulrik Hvidman, and Hans Henrik Sievertsen. 2024. "Grades and Employer Learning." *Journal of Labor Economics* 42 (3): 659-682.
- Harvard Office of Undergraduate Education. 2025. "Report on Grading." Harvard University, October.
- Hausman, Naomi, Oren Rigbi, and Sarit Weisburd. 2025. "Generative AI's Impact on Student Achievement and Implications for Worker Productivity." SSRN Working Paper 5393516.

- Hendrycks, Dan, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. "Measuring Massive Multitask Language Understanding." *International Conference on Learning Representations*.
- Hendrycks, Dan, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. "Measuring Mathematical Problem Solving with the MATH Dataset." *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*.
- OpenAI. 2022. "Introducing ChatGPT." OpenAI announcement, November 30, 2022.
- OpenAI. 2024a. "simple-evals." GitHub repository, github.com/openai/simple-evals.
- OpenAI. 2024b. "Hello GPT-4o." OpenAI announcement, May 13, 2024.
- Ostrovsky, Michael, and Michael Schwarz. 2010. "Information Disclosure and Unraveling in Matching Markets." *American Economic Journal: Microeconomics* 2 (2): 34-63.
- Pu, Zhizhong, Martin Abel, and Jeffrey P. Carpenter. 2026. "Grade Inflation and the Interpretation of Labor Market Signals." IZA Discussion Paper 18654.
- Reporting Texas. 2021. "UT Plans for an In-Person Fall Semester." May 6, 2021. <https://reportingtexas.com/ut-plans-for-an-in-person-fall-semester>
- Rojstaczer, Stuart, and Christopher Healy. 2012. "Where A Is Ordinary: The Evolution of American College and University Grading, 1940-2009." *Teachers College Record* 114 (7): 1-23.
- Roth, Jonathan. 2022. "Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends." *American Economic Review: Insights* 4 (3): 305-322.
- Scarfe, Peter, Kelly Watcham, Alasdair Clarke, and Etienne Roesch. 2024. "A Real-World Test of Artificial Intelligence Infiltration of a University Examinations System: A 'Turing Test' Case Study." *PLOS ONE* 19 (6): e0305354.
- Spence, Michael. 1973. "Job Market Signaling." *Quarterly Journal of Economics* 87 (3): 355-374.
- Strömberg, David, Victor Lei, and Yanhui Wu. 2026. "The Generative AI Learning Penalty: Evidence from Chinese Secondary Education." CEPR Discussion Paper 21577.
- Texas Legislature. 2009. H.B. No. 2504, 81st Legislature, Regular Session, enrolled version. <https://capitol.texas.gov/tlodocs/81R/billtext/html/HB02504F.htm>
- Texas Tribune. 2020. "University of Texas at Austin to Offer 2,100 Classes Online, Limit Classrooms to 40% Capacity." June 3, 2020. <https://www.texastribune.org/2020/06/03/texas-ut-austin-fall-online/>
- UT Austin. 2025a. "UT Sets All-Time Highs for Enrollment and Student Performance." UT News, September 18, 2025. <https://news.utexas.edu/2025/09/18/ut-sets-all-time-highs-for-enrollment-and-student-performance/>
- UT Austin. 2025b. "Computation of the Grade Point Average." General Information Catalog, 2025-26. <https://catalog.utexas.edu/archive/2025-26/general-information/academic-policies-and-procedures/computation-of-the-grade-point-average/computation-of-the-grade-point-average.pdf>

UT Austin. 2026. "Generative AI in Teaching and Learning: Policies." Last updated May 19, 2026. <https://fisd.utexas.edu/generative-ai-teaching-and-learning-policies>

UT Austin, EVPP Analytics and Reporting. n.d. "UT Course Grade Distributions." Accessed October 3, 2026. <https://reports.utexas.edu/spotlight-data/ut-course-grade-distributions>

Zumel Dumlao, James M., Meng Wang, Zhonghan Xie, Junyao Hu, Ivan Bar, George Chaney III, Henry Gold, and Misha Teplitskiy. 2026. "Generative AI Availability, Grades, and Student Satisfaction at a Large University." arXiv 2607.21534.

Tables

Table 1. Grade distribution by year, undergraduate sections, fall and spring semesters, 2015 to 2025

Year	Sections	Letter-grade observations	Mean grade points	Mean within-section SD	A (%)	A- or above (%)	B+ or above (%)	F (%)
2015	9,363	294,800	3.24	0.72	34.3	49.1	60.7	1.7
2016	9,292	298,075	3.27	0.71	35.7	51.0	62.5	1.7
2017	9,221	303,953	3.31	0.69	37.5	53.2	64.6	1.5
2018	9,462	309,431	3.33	0.69	39.3	54.9	66.2	1.5
2019	9,357	306,557	3.35	0.68	40.6	56.4	67.5	1.5
2020	8,773	271,754	3.61	0.53	57.2	72.8	81.5	0.9
2021	9,034	282,592	3.51	0.64	53.1	68.2	77.3	1.7
2022	9,205	311,642	3.44	0.70	49.8	64.3	73.4	1.8
2023	9,171	320,595	3.46	0.67	50.8	65.2	74.2	1.5
2024	9,048	325,572	3.52	0.62	54.3	68.5	77.1	1.2
2025	8,880	326,051	3.55	0.61	56.1	70.1	78.6	1.1

Notes: The sample includes all undergraduate sections (course numbers with last two digits from 01 to 79, excluding law, pharmacy, and medical courses) in fall and spring semesters. Letter grades are pooled across sections within each year, so that each grade counts once. Grade points follow the UT Austin scale (A = 4.00, A- = 3.67, B+ = 3.33, B = 3.00, B- = 2.67, C+ = 2.33, C = 2.00, C- = 1.67, D+ = 1.33, D = 1.00, D- = 0.67, F = 0.00). Symbols that do not enter the GPA (for example CR, NC, Q, W) are excluded. The standard deviation of grade points is computed within each section and averaged across sections, weighting by the number of letter grades.

Table 2. Downloaded syllabi by year, fall and spring semesters, 2015 to 2025

Year	Syllabi	Distinct instructor names	Subject areas	Lower-division (%)	Upper-division (%)	Graduate or professional (%)
2015	11,483	2,946	168	43.0	50.0	7.0
2016	11,430	2,956	167	44.6	50.9	4.5
2017	11,064	2,920	171	44.2	51.1	4.7
2018	11,389	3,022	173	44.2	50.0	5.7
2019	11,389	3,043	176	42.5	49.8	7.7
2020	11,905	3,000	180	41.5	50.5	8.0
2021	11,982	3,090	182	41.9	50.0	8.1
2022	12,289	3,191	182	40.7	49.6	9.7
2023	12,309	3,191	178	41.0	50.3	8.6
2024	12,461	3,201	184	40.4	50.6	9.0
2025	13,314	3,303	202	39.7	50.1	10.2

Notes: The table counts all syllabi downloaded from the UT Austin course documents site for fall and spring semesters; summer terms are excluded. Instructors are counted by full name. Subject areas are course prefixes. Course levels follow the last two digits of the course number (01 to 19 lower-division, 20 to 79 upper-division, 80 to 99 graduate); law, pharmacy, and medical courses are counted as professional.

Table 3. Consistent-course panels compared with all sections, fall and spring semesters

	All sections, 2015- 2025	Consistent courses, 2015-2025	Consistent courses, 2018-2025
Courses	7,081	698	1,027
offered in every fall		527	774
offered in every spring		459	641
offered in both		288	388
Sections	131,015	41,616	38,081
Subject areas	226	116	129
Distinct instructor names	8,092	3,690	3,740
Lower-division sections (%)	42.1	58.5	55.6
Upper-division sections (%)	50.3	41.3	42.5
Graduate or professional sections (%)	7.7	0.2	1.8
Students per section, mean	37.5	44.7	43.6
Mean grade points	3.41	3.32	3.40
Mean within-section SD	0.67	0.73	0.69
A (%)	45.9	41.1	46.4
A- or above (%)	60.9	55.6	61.0

Notes: All sections are all fall and spring sections with a downloaded syllabus from 2015 to 2025. Courses are counted by prefix and number for all sections, and by prefix, number, and title for consistent courses. A consistent course is offered in every fall, or in every spring, of the window, and only the sections of that semester are kept. Instructors are counted by full name. Course levels follow the last two digits of the course number; law, pharmacy, and medical courses are counted as professional. Grade rows use undergraduate sections with grade data; letter grades are pooled, and the standard deviation of grade points is computed within each section and averaged with weights equal to the number of letter grades.

Table 4. Instructors of consistent courses in 2025, fall and spring semesters

	Consistent courses, 2015-2025	Consistent courses, 2018-2025
Instructors in 2025	1,161	1,524
Highest degree: doctorate (%)	61.5	58.5
Highest degree: master's (%)	23.9	27.0
Years at UT Austin, mean	12.6	12.6
Years since highest degree, mean	16.0	16.0
Tenured or tenure-track faculty (%)	33.1	32.7
Teaching faculty (%)	36.2	35.4
Graduate student instructors (%)	10.7	11.7
Other roles: postdoctoral, staff, visiting (%)	3.8	3.6
Role not classified (%)	16.3	16.5

Notes: One row per instructor who taught a section of a consistent course in 2025. Degree, years at UT Austin, and role come from the current CV of the instructor. Teaching faculty include lecturers and professors of practice.

Table 5. Consistent-instructor panel, 2018 to 2025

	Fall track	Spring track	Combined panel
Instructors	403	335	529
Sections	10,978	9,607	20,585
Courses taught	1,414	1,303	2,150
Subject areas	145	128	152
Sections per instructor-track pair per year, mean	3.41	3.58	3.49
Courses per instructor-track pair per year, mean	1.97	2.02	1.99
Same course in all 8 years (% of instructor-track pairs)	44.2	44.5	44.3
Sections in a consistent course, 2018-2025 (%)	65.8	61.1	63.6
Lower-division sections (%)	44.5	40.8	42.8
Upper-division sections (%)	49.6	54.6	51.9
Graduate or professional sections (%)	5.9	4.6	5.3
Mean grade points	3.41	3.47	3.44
A (%)	46.4	50.5	48.2

Notes: An instructor is on the fall (spring) track if the instructor has at least one syllabus in every fall (spring) semester from 2018 to 2025, in any course. The combined panel counts each of the 529 instructors once; 209 instructors are on both tracks. Per-year averages and the share who taught the same course in all eight years are computed over instructor-track pairs, so an instructor on both tracks contributes two pairs. Grade rows use undergraduate sections with grade data, with letter grades pooled.

Table 6. AI capability by task type, binary rule (free ChatGPT tier, 2024-25)

Task type	Examples	AIcap
Written prose	essay, paper, memo, report, reflection, literature review	1
Objective items	multiple choice, short answer, quiz	1
Code or data	programming, scripts, data analysis, notebooks	1
Quantitative problem set	math, statistics or economics problem sets, derivations, proofs	0
Unspecified exam	exam with no format given	0
Unspecified other	graded work with no format given	0
Oral or performance	presentation, speech, recital, in-class discussion	0
Studio, lab or physical	studio work, physical lab, fieldwork, clinical practice	0
Participation	attendance, participation, peer review	0

Notes: A task type is coded as AI-doable (AIcap = 1) if GPT-4o, the most capable model on the free tier of ChatGPT in the 2024-25 academic year, scored at least 70 percent, a passing grade in most US courses, on the public benchmark closest to an undergraduate assignment of that type. Written prose has no benchmark with a pass rate; its coding rests on course-grading evidence. Types with no clear task are set to zero. Oral work, studio and lab work, and participation are zero by construction, because the student produces the graded work in person. Benchmark scores are given in Table A1, and scores for each task type and year are in Table A2, both in the Online Appendix.

Table 7. Unproctored share from a research assistant, Gemini 3.7 Flash and GLM-5.3, 500-syllabus benchmark, 2019 and 2022

Statistic	Research assistant	Gemini 3.7 Flash	GLM-5.3
Mean unproctored share	0.437	0.457	0.445
Mean absolute difference from research assistant		0.079	0.083
Identical to research assistant (%)		65.2	63.1
Within 0.10 of research assistant (%)		78.0	76.3
Correlation with research assistant		0.820	0.808

Notes: The sample is the 477 syllabi, out of the 500 coded by hand, that all three coders could score; the others lack weights or could not be read, and GLM-5.3 failed on one syllabus. The unproctored share is the share of the course grade from components coded as unproctored; proctored components and components with an unclear format count as zero, as in the main measure. "Identical" means that the two shares are equal.

Table 8. Descriptive statistics of AI capability, unproctored share and AI exposure, 2022 and 2019 syllabi, consistent courses 2018-2025

Measure	Courses	Mean	SD	25th percentile	Median	75th percentile	Zero (%)
<i>Panel A. 2022 syllabi</i>							
AI capability	997	0.387	0.310	0.075	0.370	0.630	19.3
Unproctored share	997	0.505	0.285	0.265	0.500	0.750	4.8
AI exposure	997	0.297	0.276	0.014	0.250	0.500	24.4
<i>Panel B. 2019 syllabi</i>							
AI capability	990	0.378	0.316	0.050	0.350	0.650	21.9
Unproctored share	990	0.370	0.258	0.163	0.305	0.550	8.1
AI exposure	990	0.230	0.242	0.000	0.166	0.394	31.5

Notes: One observation per course. Each measure is a share of the course grade: AI capability is the share from AI-doable task types (written prose, objective items, and code or data), the unproctored share is the share from work done without a proctor, and AI exposure is the share from work that is both. The value of a course is the mean over its sections in the base year, weighted by the number of students. "Zero" is the share of courses with a value of exactly zero. The sample is the consistent courses of the 2018-2025 panel with at least one coded syllabus with grade weights in the base year.

Table 9. Mean AI capability, unproctored share and AI exposure by course group, 2022 syllabi, consistent courses 2018-2025

Group	Courses	AI capability	Unproctored share	AI exposure
Course level: lower division	393	0.414	0.514	0.313
Course level: upper division	558	0.360	0.488	0.272
Course level: graduate	46	0.484	0.635	0.466
Students per section: fewer than 25	399	0.391	0.516	0.316
Students per section: 25 to 99	351	0.358	0.471	0.241
Students per section: 100 or more	78	0.511	0.603	0.384
Field: humanities and arts	293	0.456	0.571	0.366
Field: social sciences	65	0.553	0.545	0.379
Field: natural sciences and mathematics	143	0.244	0.437	0.178
Field: engineering	125	0.190	0.417	0.154
Field: professional schools	371	0.425	0.502	0.323
All courses	997	0.387	0.505	0.297

Notes: Means over courses; the value of each course is the mean over its 2022 sections, weighted by the number of students. Course levels follow the course number, as in Section 2. Students per section is the number of students in the course in 2022 divided by its number of sections; 169 courses with no enrollment record in 2022 are not included in these three rows. Fields group subject areas as follows: humanities and arts (languages, literature, history, philosophy, area studies, fine arts), social sciences (anthropology, economics, government, geography, psychology, sociology, and related programs), natural sciences and mathematics (including computer science and statistics), engineering, and professional schools (business, communication, education, kinesiology and health, nursing, social work, architecture, and military science). The course prefixes in each field are listed in Table A4 of the Online Appendix.

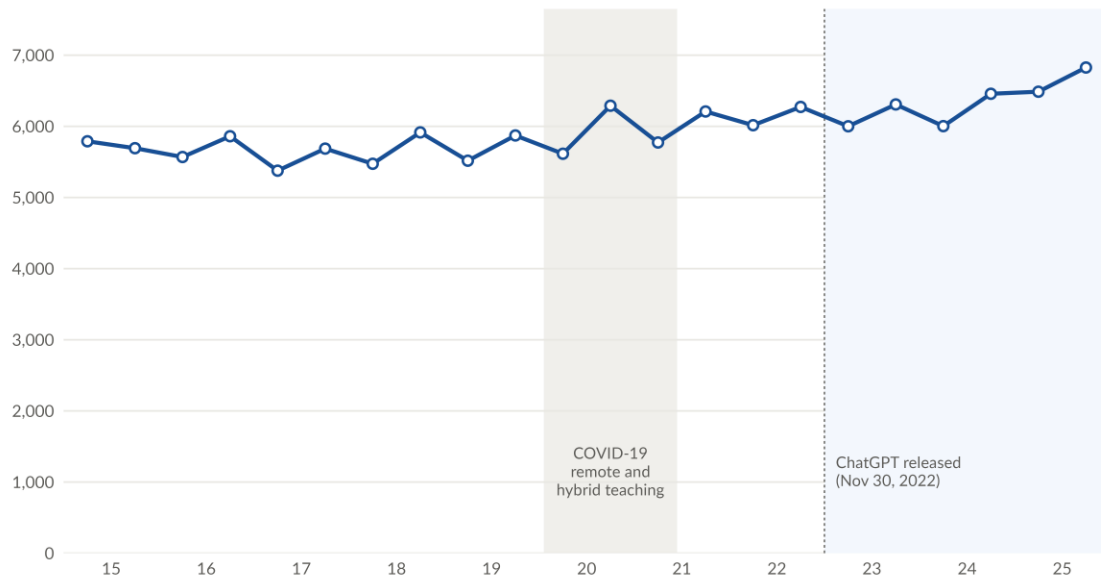
Table 10. Difference-in-differences estimates, AI capability and AI exposure (2022 syllabi), balanced panel of consistent courses, 2018-2025, unweighted

Treatment Semesters	(1) AI capability Fall	(2) AI capability Fall and spring	(3) AI exposure Fall	(4) AI exposure Fall and spring
Panel A. Shares of letter grades				
Share of A	0.085*** (0.017) [0.446]	0.067*** (0.013) [0.463]	0.100*** (0.020) [0.446]	0.080*** (0.015) [0.463]
Share of A-	-0.034*** (0.009) [0.162]	-0.020*** (0.006) [0.158]	-0.040*** (0.011) [0.162]	-0.029*** (0.008) [0.158]
Share of B+	-0.016*** (0.006) [0.104]	-0.005 (0.005) [0.102]	-0.025*** (0.007) [0.104]	-0.013** (0.005) [0.102]
Panel B. Cumulative shares				
A- or above	0.051*** (0.015) [0.608]	0.047*** (0.012) [0.622]	0.061*** (0.018) [0.608]	0.051*** (0.014) [0.622]
B+ or above	0.035** (0.014) [0.712]	0.042*** (0.011) [0.724]	0.036** (0.017) [0.712]	0.038*** (0.013) [0.724]
B or above	0.020** (0.010) [0.829]	0.027*** (0.008) [0.838]	0.023* (0.012) [0.829]	0.028*** (0.009) [0.838]
Panel C. Mean and dispersion of grade points				
Mean grade points	0.067*** (0.025) [3.411]	0.071*** (0.020) [3.434]	0.078** (0.031) [3.411]	0.076*** (0.024) [3.434]
Mean within-section SD	-0.036* (0.020) [0.664]	-0.049*** (0.016) [0.649]	-0.033 (0.027) [0.664]	-0.050** (0.020) [0.649]
Panel D. Low grades				
Share of F	0.001 (0.002) [0.015]	0.002 (0.002) [0.016]	0.000 (0.003) [0.015]	0.001 (0.002) [0.016]
Share of D or F	-0.001 (0.004) [0.032]	-0.003 (0.003) [0.031]	-0.001 (0.005) [0.032]	-0.003 (0.004) [0.031]
Course-semesters	3,944	7,272	3,944	7,272
Courses	493	634	493	634
Letter-grade observations	693,259	1,249,177	693,259	1,249,177

Notes: Each coefficient is the estimate of β in equation (3) from a separate regression, in which the treatment named in the column, measured from the 2022 syllabi of the same term, is interacted with an indicator for the semesters from spring 2023 on (from fall 2023 on for fall tracks). AI capability is the share of the course grade from task types that AI can do, and AI exposure is the share from work that AI can do and that is completed without a proctor. The unit is a course in a semester, with all sections combined, and each observation has equal weight. The sample is the balanced panel of consistent undergraduate courses, in which every track has letter grades in all eight years from 2018 to 2025. Columns (1) and (3) include track and year fixed effects, and columns (2) and (4) include track and semester fixed effects. The interquartile range of the treatment is 0.663, 0.617, 0.450 and 0.471 in columns (1) to (4). Standard errors, clustered by course, are in parentheses, and the mean of the outcome before 2023 is in brackets. Shares are between 0 and 1, so 0.010 equals 1 percentage point; mean grade points and the mean within-section SD are in grade points on the UT Austin scale, and the share of D or F counts D+, D, D- and F grades. Letter-grade observations count student-course records with a letter grade. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

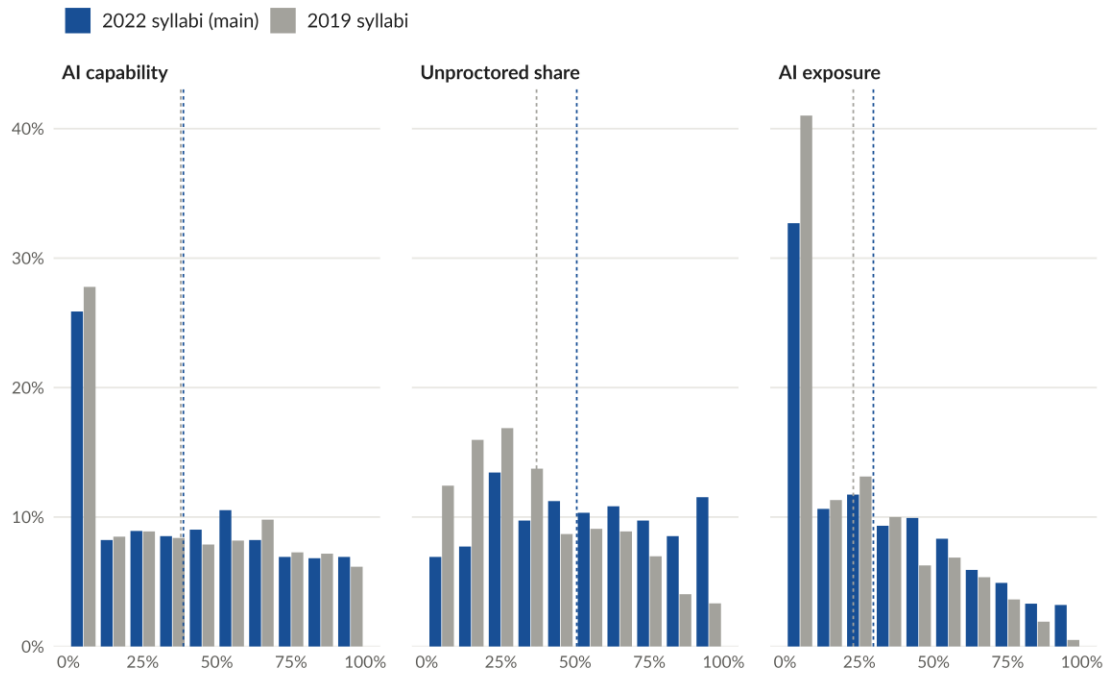
Figures

Figure 1. Syllabi downloaded per semester, fall and spring, 2015 to 2025



Notes: The figure shows the number of syllabi downloaded from the UT Austin course documents site in each semester. Each point is one semester (spring, then fall, within each year); summer terms are not shown. The grey area marks spring 2020 to spring 2021 (remote and hybrid teaching). The blue area marks the period after the release of ChatGPT on November 30, 2022.

Figure 2. Distribution of AI capability, unproctored share and AI exposure across courses, consistent courses, 2022 and 2019 syllabi



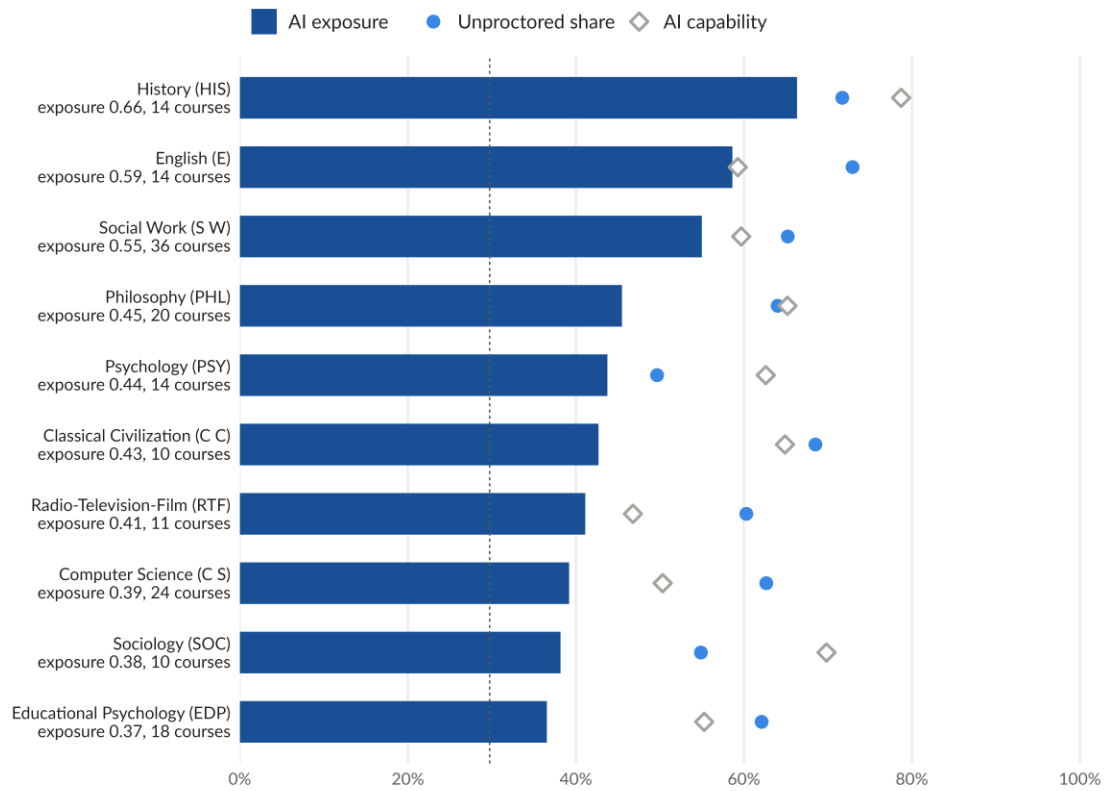
Notes: Each panel shows the share of courses in each 10-point bin of the measure; dashed lines are the means. One observation per course: 997 courses in 2022 and 990 in 2019.

Figure 3. AI capability and unproctored share across courses, consistent courses, 2022 syllabi



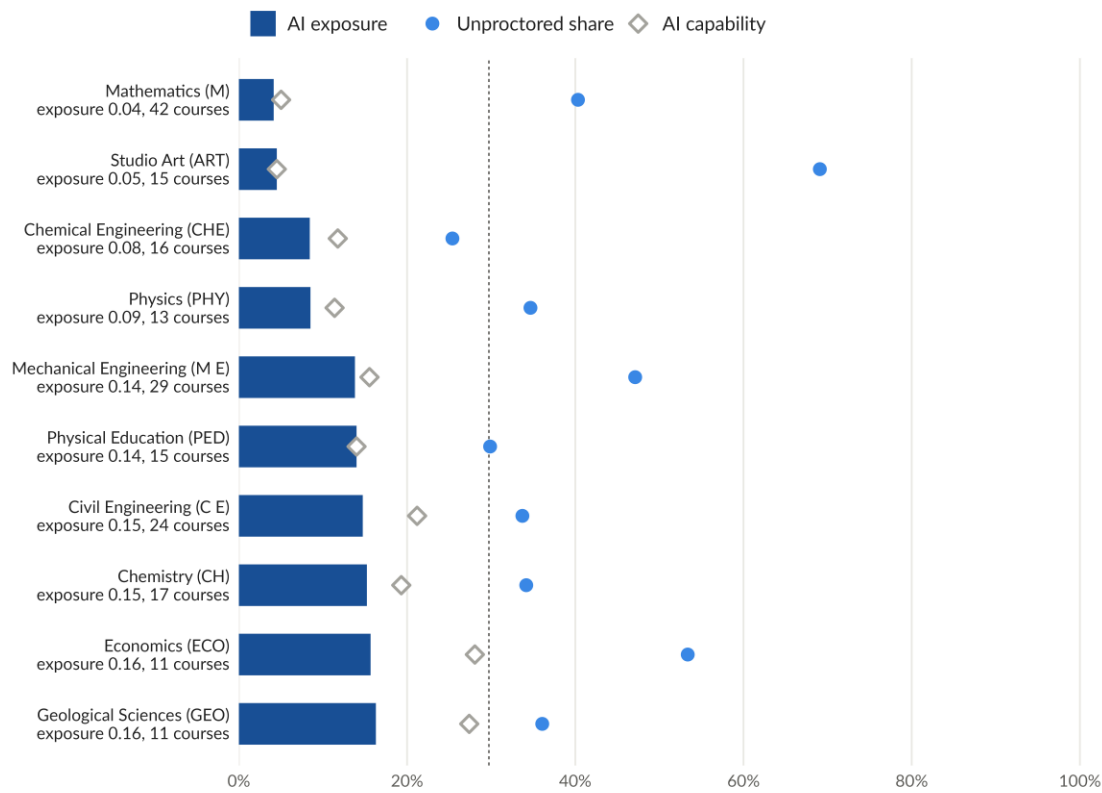
Notes: One point per course (997 courses). The colour of each point shows its AI exposure. Points are moved slightly at random so that courses with equal values remain visible.

Figure 4. The ten most exposed departments, consistent courses, 2022 syllabi



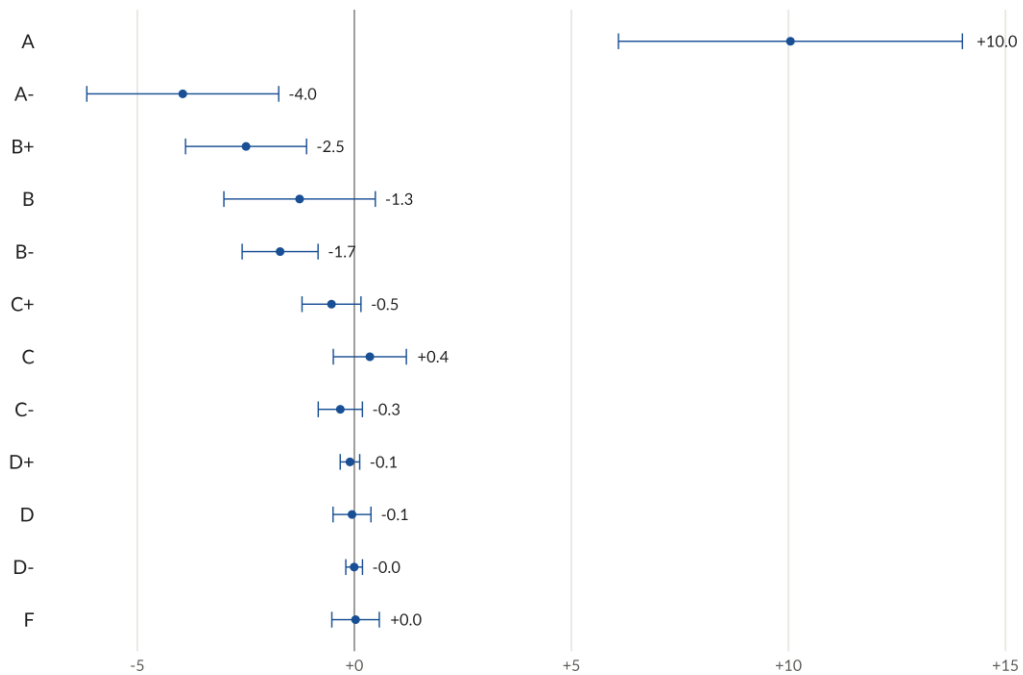
Notes: Department means over courses, for departments with at least 10 measured consistent courses (34 departments). Bars show AI exposure, dots the unproctored share, and diamonds AI capability. The dashed line is the mean AI exposure of all 997 courses (0.30).

Figure 5. The ten least exposed departments, consistent courses, 2022 syllabi



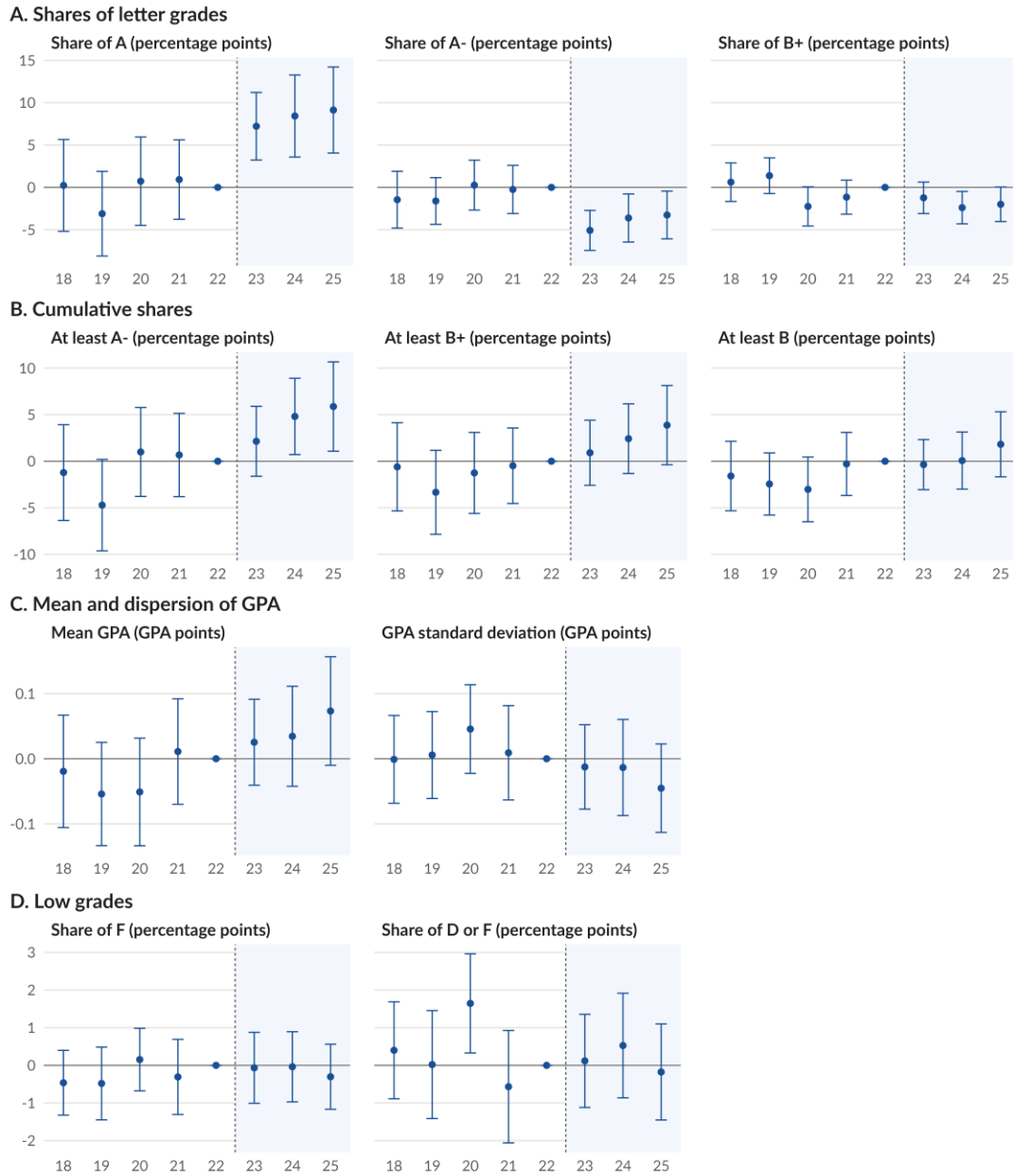
Notes: As in Figure 4.

Figure 6. AI exposure and the share of each letter grade, fall tracks, 2022 syllabi, 2018-2025



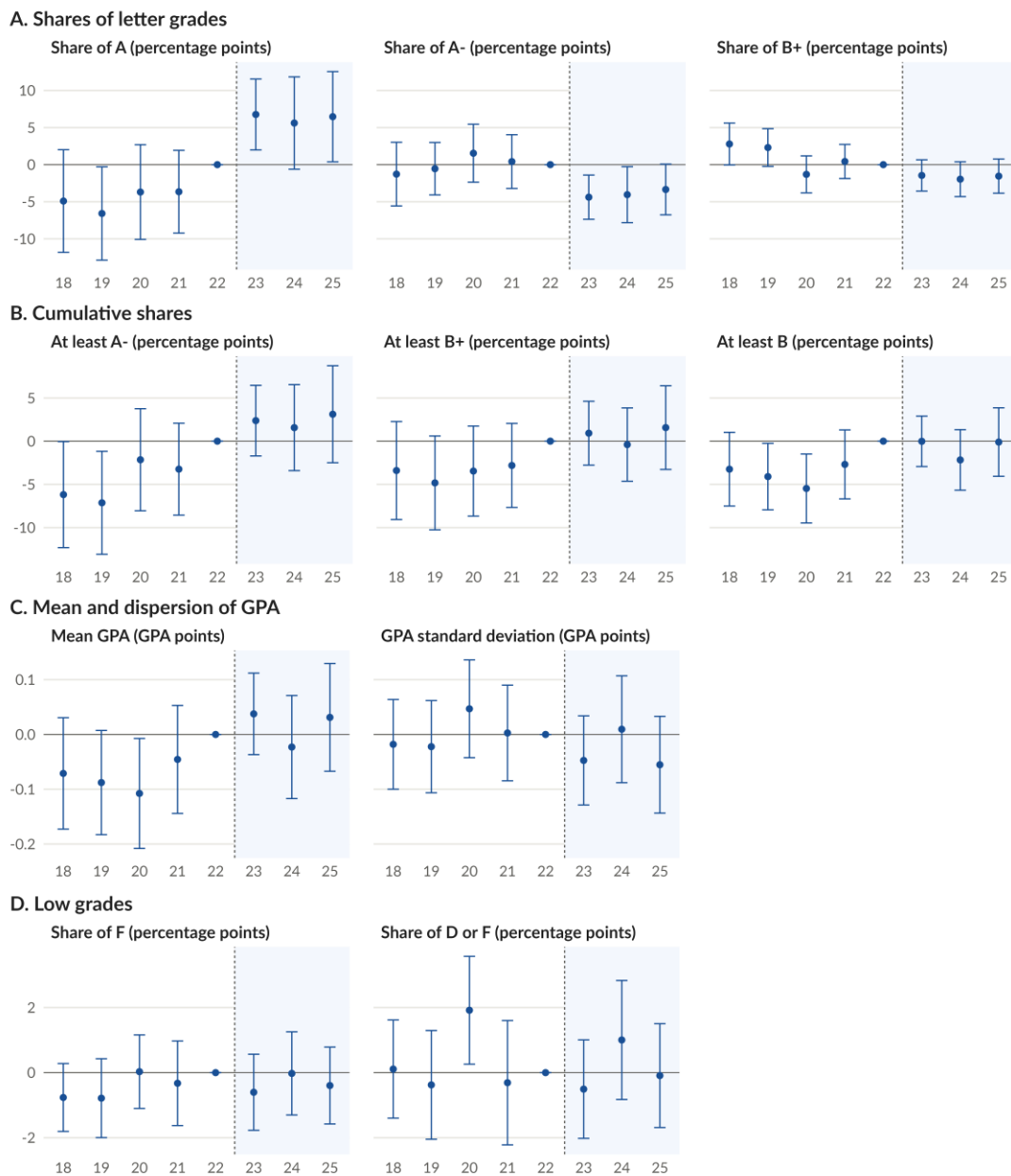
Notes: Each row reports the estimate of β in equation (3) from a separate regression in which the outcome is the share of that letter grade among all letter grades in the course and semester, in percentage points, for a one-unit change in AI exposure. AI exposure is measured from the fall 2022 syllabi, and the sample is the 493 fall tracks of the balanced panel of consistent courses, 2018-2025, with track and year fixed effects and equal weights. Lines are 95 percent confidence intervals based on standard errors clustered by course.

Figure 7. Event study, AI capability (2022 syllabi), fall tracks, 2018-2025



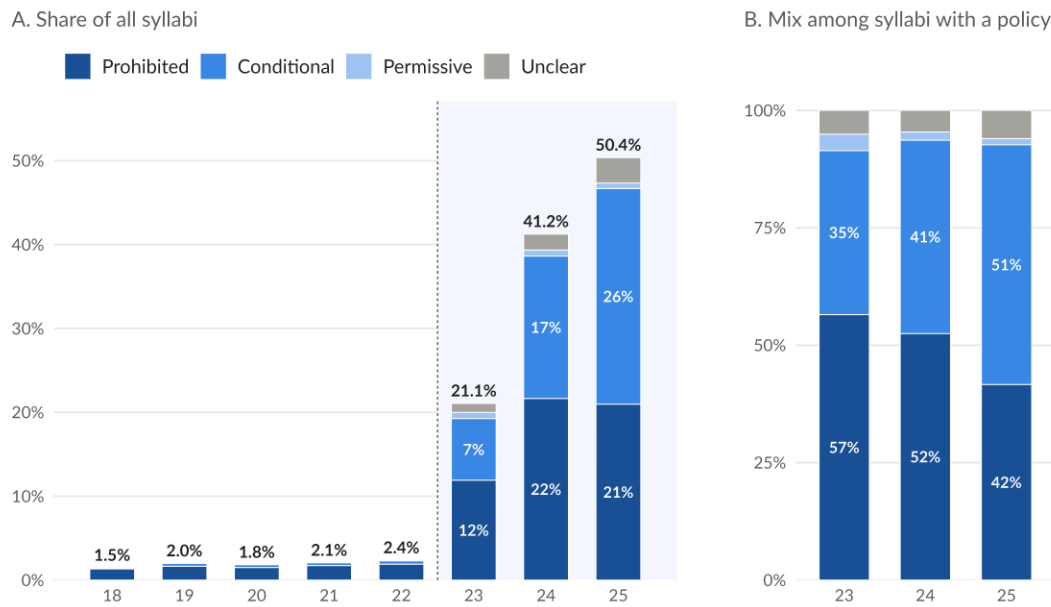
Notes: Each point is the estimate of β_k in equation (4), the change in the outcome relative to 2022 for a one-unit change in AI capability, measured from the fall 2022 syllabi. Shares are in percentage points, and mean grade points and the mean within-section SD are in grade points; the share of D or F counts D+, D, D- and F grades. Panels in the same row share one y scale. The shaded area marks the years after the release of ChatGPT in November 2022. The sample is the 493 fall tracks of the balanced panel of consistent courses, 2018-2025, with track and year fixed effects and equal weights. Lines are 95 percent confidence intervals based on standard errors clustered by course.

Figure 8. Event study, AI exposure (2022 syllabi), fall tracks, 2018-2025



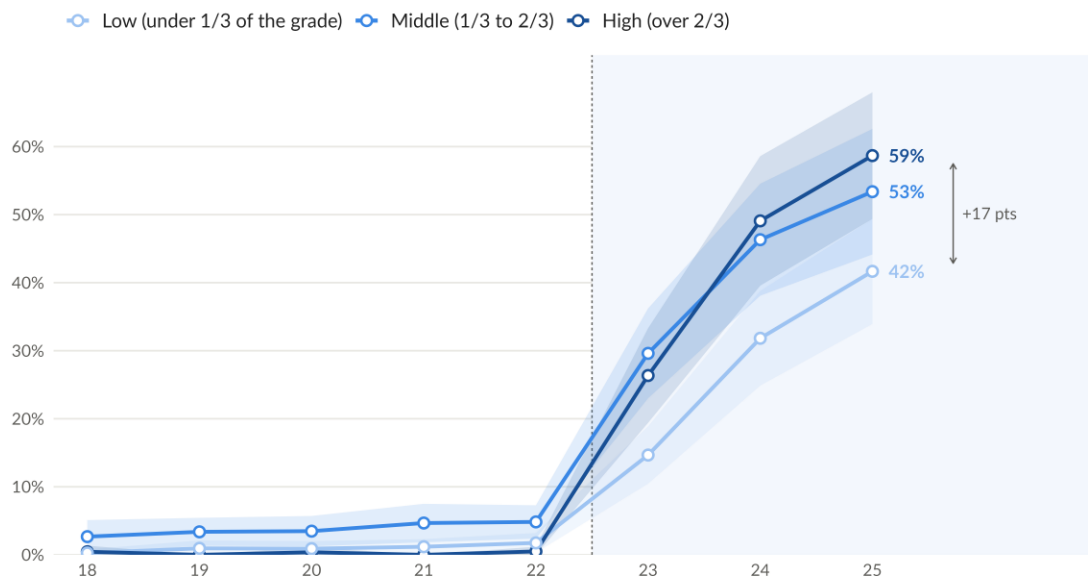
Notes: As in Figure 7, with AI exposure, the share of the course grade from work that AI can do and that is completed without a proctor, in place of AI capability.

Figure 9. AI policy in syllabi by year and type, panel of consistent instructors, 2018-2025



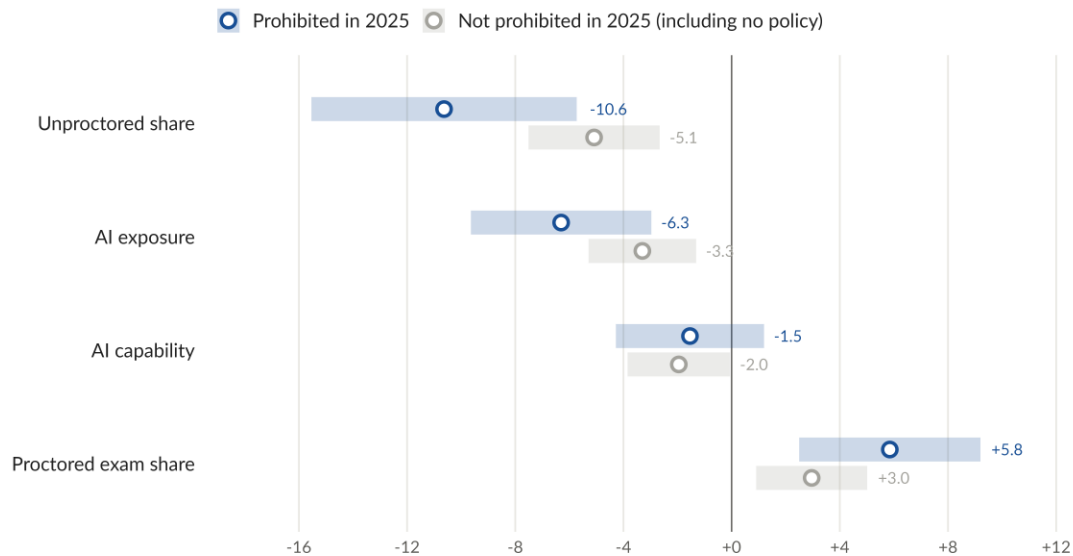
Notes: In Panel A, each bar is the share of all syllabi in the year with an AI policy of each type, and the number above the bar is the share with any policy. In Panel B, each bar splits the syllabi with a policy in 2023, 2024 and 2025 (315, 629 and 764 syllabi) by type. A syllabus for one course in one term counts once, which gives 1,430 to 1,525 syllabi per year, and each year includes the spring and fall semesters. Before 2023, the codes are errors, since no syllabus coded with a policy contains a word related to AI. The shaded area marks the years after the arrival of GenAI in November 2022.

Figure 10. Share of syllabi with an AI policy by AI capability in 2022, panel of consistent instructors, 2018-2025



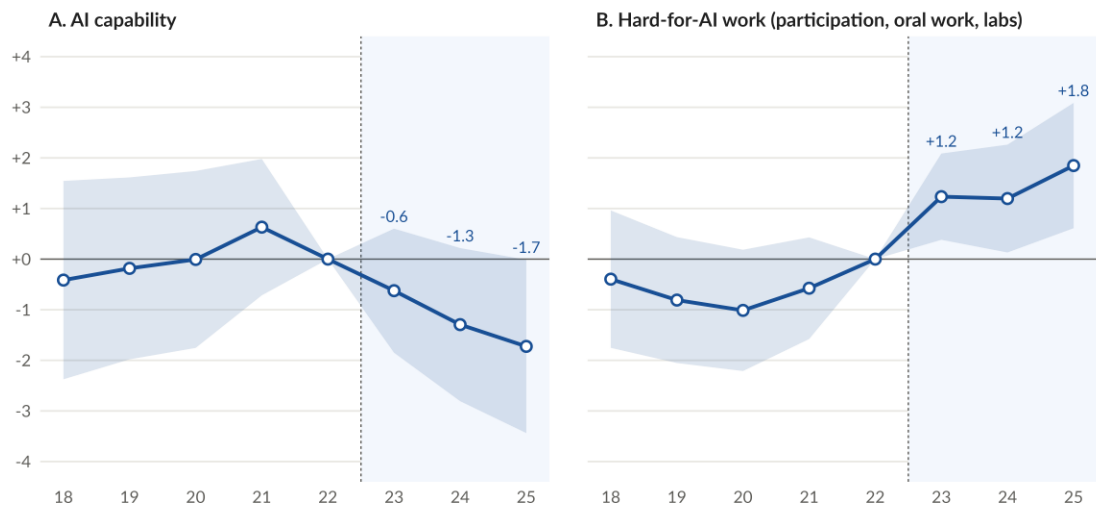
Notes: Each line is the share of syllabi in the group with any AI policy (prohibited, conditional, permissive or unclear) in each year. AI capability is the share of the course grade from written prose, objective items, and code or data, measured as the mean over the 2022 syllabi of the same instructor and course. Low AI capability means that less than one third of the grade comes from these task types, middle means one third to two thirds, and high means more than two thirds. The sample is the 8,657 syllabi of the 1,210 pairs of instructor and course with at least one syllabus in 2022 (483 low, 372 middle and 355 high); a syllabus for one course in one term counts once, and each year includes the spring and fall semesters. The shaded bands are 95 percent confidence intervals, with standard errors clustered by instructor. Before 2023, the codes are errors, since no syllabus coded with a policy contains a word related to AI. The shaded area marks the years after the arrival of GenAI in November 2022.

Figure 11. Change in syllabus design from 2022 to 2025 by AI policy in 2025, panel of consistent instructors



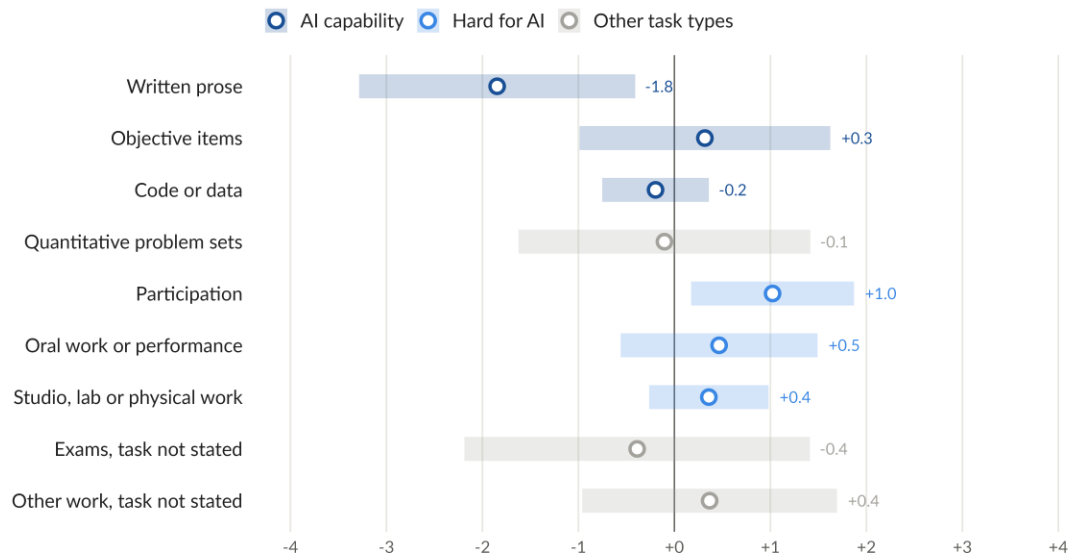
Notes: The sample is the 742 pairs of instructor and course with at least one syllabus in both 2022 and 2025. A pair is in the prohibited group (159 pairs) when its 2025 syllabi ban AI, and in the other group (583 pairs) when they have a conditional, permissive or unclear policy or no policy. Each point is the mean change in the share of the course grade from 2022 to 2025, in percentage points, where each share is first averaged over the syllabi of the pair in the year. AI capability is the share of the grade from written prose, objective items, and code or data, and AI exposure is the share from such work done without a proctor; components with an unclear format and work done live online count as proctored. The proctored exam share is the share of the grade from exams, midterms, quizzes and tests that are proctored. Shaded bars are 95 percent confidence intervals for the change in each group, with standard errors clustered by instructor. The intervals show the uncertainty of each group's change on its own, so two intervals can overlap even when the gap between the groups is statistically significant: the standard error of the gap in the unproctored share is 2.8 percentage points, smaller than the sum of the standard errors of the two groups (2.5 and 1.2 percentage points).

Figure 12. Change in grade weights relative to 2022, same instructor and course, panel of consistent instructors, 2018-2025



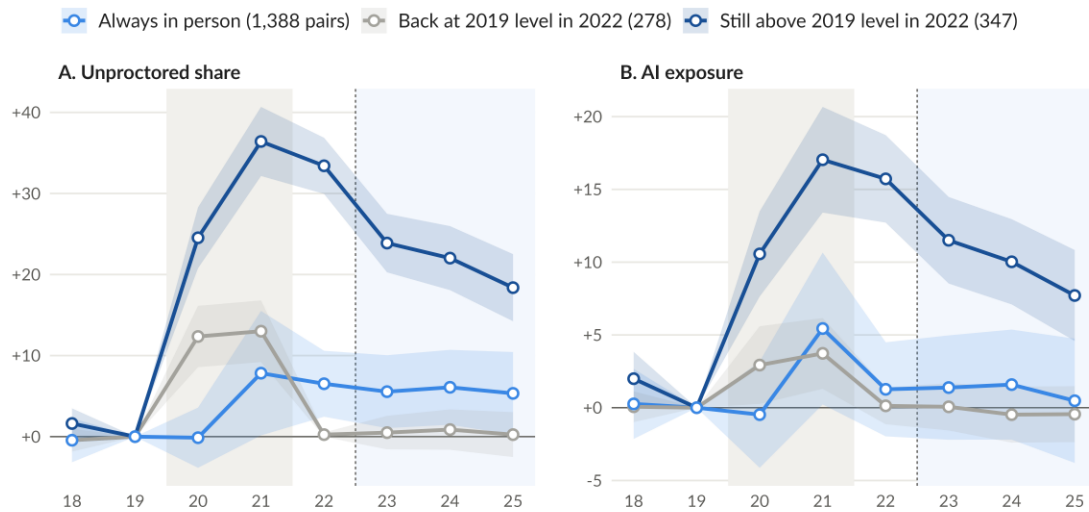
Notes: Each point is the change in the share of the course grade relative to 2022, in percentage points, from an event study with fixed effects for each pair of instructor and course and an indicator for each year, with 2022 as the omitted year. AI capability is the share of the grade from written prose, objective items, and code or data, and hard-for-AI work is the share from participation, oral work or performance, and studio, lab or physical work. The sample is 9,604 observations of 2,784 pairs of instructor and course (528 instructors), where each observation is the mean over the syllabi of a pair in a year, and shares use the components in the main weight unit of each syllabus. Shaded bands are 95 percent confidence intervals, with standard errors clustered by instructor. The blue area marks the years after the arrival of GenAI in November 2022.

Figure 13. Change in grade weights by task type from 2022 to 2025, same instructor and course, panel of consistent instructors



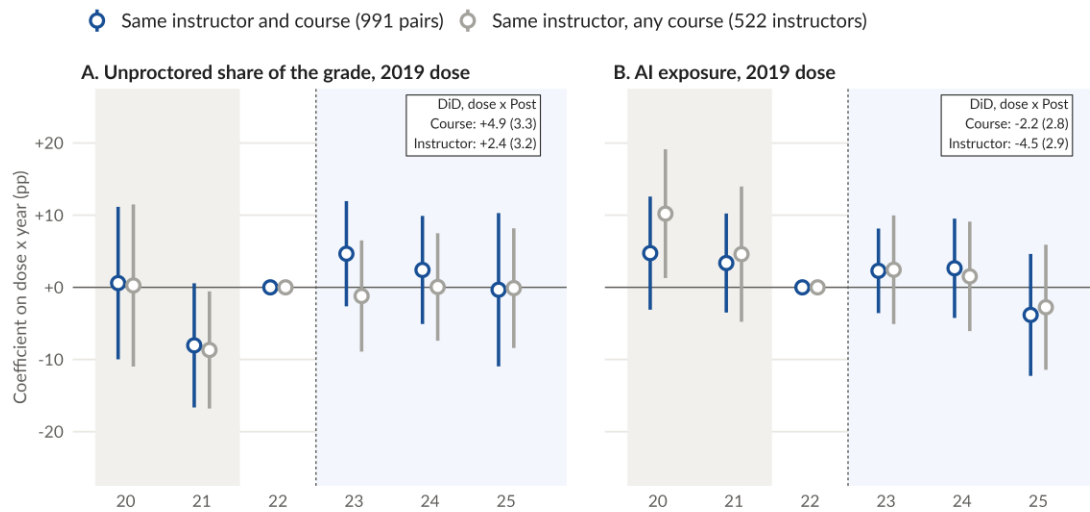
Notes: Each point is the 2025 coefficient of a separate event study, as in Figure 12, for the share of the course grade from one task type, in percentage points. Dark blue marks the task types counted in AI capability, light blue the task types that are hard for AI, and grey the other task types, including components whose task is not stated in the syllabus. Shaded bars are 95 percent confidence intervals, with standard errors clustered by instructor.

Figure 14. Change in proctoring relative to 2019 by COVID group, same instructor and course, panel of consistent instructors, 2018-2025



Notes: Each point is the change in the share of the course grade relative to 2019, in percentage points, from an event study with fixed effects for each pair of instructor and course and an indicator for each year, with 2019 as the omitted year, estimated separately for each group. Always in person: 1,388 pairs whose syllabi report in-person teaching in every observed year. Back at 2019 level: 278 pairs whose unproctored share in 2022 is within 5 percentage points of their 2019 share. Still above 2019 level: 347 pairs whose unproctored share in 2022 is more than 5 percentage points above their 2019 share. AI exposure is the share of the grade from written prose, objective items, and code or data that is done without a proctor, and components with an unclear format and work done live online count as proctored. Shaded bands are 95 percent confidence intervals, with standard errors clustered by instructor. The grey area marks the COVID years 2020 and 2021, and the blue area marks the years after the arrival of GenAI in November 2022.

Figure 15. Change in proctoring by its 2019 level, two comparisons, panel of consistent instructors, 2020-2025



Notes: Each point is the coefficient on the 2019 value of the outcome times an indicator for the year, in percentage points for a one-unit change in the 2019 value, from a regression with year fixed effects and 2022 as the omitted year, using only the years 2020 to 2025. Blue: fixed effects for each pair of instructor and course, with the pair's 2019 value as the dose (991 pairs with a 2019 syllabus, 4,070 pair-years). Grey: fixed effects for each instructor, where each year is the mean over all syllabi of the instructor and the dose is the instructor's 2019 value (522 instructors, 3,086 instructor-years). The boxes report the difference-in-differences coefficient on the dose times an indicator for the years from 2023 on, with standard errors in parentheses. Lines are 95 percent confidence intervals, with standard errors clustered by instructor. The grey area marks the COVID years 2020 and 2021, and the blue area marks the years after the arrival of GenAI in November 2022.