

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

Reza Tavakoli Jaghargh

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

Online Appendix, Draft V4, October 2026

This online appendix accompanies the paper "From A-Minus to A: Generative AI and Grade Compression at the Top." Tables, figures and sections with a letter (for example, Table A1 or Section B.2) are in this appendix, while those with a number only (for example, Table 10 or Section 3) are in the paper.

Appendix A: Data and Coding Details

This appendix gives the details behind the measures used in the paper, namely the benchmarks that define AI capability, the coding of the syllabi and of the instructor CVs with large language models, and the grouping of subject areas into fields, while the prompts themselves are printed in Appendix E.

A.1 Benchmarks for AI Capability

Table A1 lists every benchmark that OpenAI reported for GPT-4o in its simple-evals repository (OpenAI 2024a), together with U-MATH and with the task type in the syllabi to which each benchmark maps and the code that the 70 percent rule gives to that type. Three of these benchmarks match a graded task type directly, namely MMLU for objective items, HumanEval for code or data and MATH for quantitative problem sets, and GPT-4o is above the 70 percent line on all three. U-MATH, which is not in the OpenAI set, is used for problem sets in place of MATH because its problems come from university courses rather than from math competitions.

The other four benchmarks test skills that do not match a graded task type in undergraduate syllabi, so they are listed but do not enter the coding. MGSM tests grade-school arithmetic and DROP tests reading comprehension, and GPT-4o is above the line on both in any case. GPQA and SimpleQA were built to be hard for models, because GPQA questions were written so that skilled non-experts with web access answer only about a third of them correctly, and SimpleQA questions were collected so that GPT-4 tended to answer them wrongly.

Table A1. Benchmarks that OpenAI reported for GPT-4o and the coding of AI capability

Benchmark	What it tests	Task type in this paper	GPT-4o score (%)	Alcap
MMLU (Hendrycks et al. 2021a)	Multiple-choice questions in 57 subjects, from high school to professional level	Objective items	88.7	1
HumanEval (Chen et al. 2021)	Python functions written from a short description	Code or data	90.2	1
MATH (Hendrycks et al. 2021b)	Short problems from math competitions	Quantitative problem sets	75.9	Not used
U-MATH (Chernyshev et al. 2024)	Free-response problems from US university math courses	Quantitative problem sets	About 50	0
DROP (Dua et al. 2019)	Reading comprehension with counting and arithmetic	Not mapped	79.8	
MGSM (Shi et al. 2022)	Grade-school math word problems in ten languages	Not mapped	90.0	
GPQA (Rein et al. 2023)	Graduate-level science questions that are hard to answer by web search	Not mapped	53.1	
SimpleQA (Wei et al. 2024)	Short questions on facts, answered without search	Not mapped	40.1	
No benchmark in the set		Written prose		1

Notes: Scores are the accuracy, in percent, of gpt-4o-2024-08-06 in the simple-evals repository (OpenAI 2024a), except DROP, which is an F1 score, and U-MATH, which is not part of the OpenAI set and is taken from Chernyshev et al. (2024). A task type is coded as AI-doable (Alcap = 1) when GPT-4o scores at least 70 percent on the benchmark closest to it. Written prose has no benchmark in the set, and it is coded as AI-doable on the basis of the course-grading evidence in Section 3.1.1. Coding quantitative problem sets as AI-doable, as the MATH score implies, is the check reported in row (2c) of Tables B1 and B2. The other versions of GPT-4o in the repository give similar scores, from 85.7 to 88.7 percent on MMLU and from 90.2 to 91.0 percent on HumanEval.

AI capability is coded for one year, 2024-25, but the scores of the free ChatGPT model changed from year to year, and Table A2 shows these scores for the three task types that have a benchmark with a pass rate. Objective items were at the 70 percent line in 2022-23 and 2023-24 and well above it from 2024-25, while code or data passed the line from 2023-24, the first full academic year after the release of ChatGPT. Quantitative problem sets passed the line on MATH from 2024-25, but on U-MATH, the benchmark closer to university work, the free model stayed below it through 2024-25 (Table A2), and problem sets reach the line on the free tier only in 2025-26, when reasoning models became free.

Table A2. Scores of the free ChatGPT model on the benchmarks for each task type, by academic year (percent)

Academic year	Free ChatGPT model	Objective items (MMLU)	Code or data (HumanEval)	Problem sets (MATH)	Problem sets (U-MATH)
2022-23	GPT-3.5	70.0	48.1	34.1	
2023-24	GPT-3.5 Turbo	70.1	74.4	34.1	
2024-25	GPT-4o	88.7	90.2	75.9	About 50
2024-25	GPT-4o mini, after the usage limit	82.0	87.2	70.2	About 43
2025-26	GPT-5 (scores of o3)	92.9	87.4	97.8	

Notes: Each score is the share of the benchmark answered correctly by the model on the free tier of ChatGPT in the fall of each academic year. Scores for 2022-23 are from the GPT-4 Technical Report (OpenAI 2023) for MMLU and HumanEval and from Gemini Team (2023) for MATH. Scores for 2023-24 are from Akter et al. (2023), and the MATH score of 2022-23 is carried forward because no later measurement of GPT-3.5 Turbo on MATH was found. Scores for 2024-25 are from the simple-evals repository (OpenAI 2024a), and U-MATH scores are from Chernyshev et al. (2024). OpenAI did not report GPT-5 on these older benchmarks, so the 2025-26 row shows the scores of o3, a reasoning model released earlier in 2025 (OpenAI 2024a). Written prose has no benchmark with a pass rate in any year.

A.2 Coding of the Syllabi

The syllabi of the course panel were coded with Gemini 3.7 Flash at the medium reasoning level, and the syllabi of the instructor panel were coded with GLM-5.3 at the high reasoning level, which was run through the language-model service of Advanced Research Computing at Virginia Tech. Both models received the same prompt and returned their output in the same format, and every request used a temperature of 0 and a fixed seed (20260812; Table E1), so that the same syllabus gives the same answer when it is read again. The output had to follow a fixed JSON schema that lists the allowed values of every field, and the measures in Section 3 were computed afterwards from this output by the same code for both models.

Each request contained one syllabus, whose text was extracted from the PDF file, while syllabi that were scanned as images were sent as page images instead. A line with the registered course number was placed before the text, so that the model could pick the grading scheme of the right course when one document covers several courses or both undergraduate and graduate students. The middle pages of long syllabi were left out and marked as omitted, and the model was told to code only the text that it could see.

The prompt has two parts, a short preamble and a rulebook of about 3,800 words, and the preamble tells the model that it is coding one university syllabus against a fixed rulebook, that it must code the document from its text alone, and that it must return one JSON object with the fields in Table A3. The rulebook first asks the model to check that the document is a syllabus, and then to find the grading breakdown, list every graded component with its weight, and leave

out bonus points and extra credit. For each component, the model records where the work is completed, whether it is proctored, and what kind of work it is, and the rulebook states that the task type must not depend on the answers about where the work is completed. The full preamble and rulebook are printed in Appendix E, and the code that sends them to the models is provided with the replication files of this paper.

Table A3. Fields returned by the model for each syllabus

Field	Level	Allowed values
Course level	Syllabus	Undergraduate, graduate, unknown, not a syllabus
Delivery mode	Syllabus	In-person, online, hybrid, unknown
Online modality	Syllabus	Asynchronous, synchronous, mixed, not applicable, unknown
Class sessions	Syllabus	Four yes-or-no items: any in-person sessions, any online sessions, student choice between the two, and a short remote start only
AI policy	Syllabus	Prohibited, conditional, permissive, unclear, no policy
Grading basis	Syllabus	Letter grades, pass or fail, optional pass or fail, unknown
Component name	Component	Text as printed in the syllabus
Weight and unit	Component	A number, in percent, points or count
Administration	Component	In person, online live, online with a named proctoring tool, online and self-paced, take-home, not applicable, unknown
Proctoring	Component	Proctored, unproctored, indeterminate
Task type	Component	Written prose, code or data, quantitative problem set, objective items, oral or performance, studio or lab, participation, unspecified (later split into unspecified exams and other unspecified components, as in Table 6)
Task evidence	Component	The words in the syllabus that support the task type, and where they appear

Notes: Each syllabus returns one value for each syllabus-level field and a list of graded components, each with the component-level fields. Proctoring follows from administration: in person, online live and online with a named proctoring tool are proctored, online and self-paced and take-home are unproctored, and indeterminate is used only when one weight covers both kinds of work, while components whose administration is not applicable or unknown are coded as proctored.

The course panel was coded in three runs in August and September 2026, which covered the 2022 syllabi, the 2019 syllabi, and the syllabi of the other years. Identical documents shared by several sections were coded only once, so the 2022 run, which covers 4,833 sections, needed about 2,590 requests and cost about 13 dollars. In this run, 77 documents (about 3 percent of the total), which cover 187 sections, returned no graded components, mostly because the syllabus gives no weights or uses specifications grading, and these sections have no measure.

A.3 Coding of the Instructor CVs

Instructor CVs were coded with Qwen3.6-27B, an open-weight model, which was run on the graphics processors of the Advanced Research Computing cluster at Virginia Tech with a temperature of 0, a fixed seed, and the reasoning mode turned off. Each request contained the text of one CV, with the surname of the instructor on file placed before the text as a hint, and the later pages of long CVs were left out and marked as omitted. The model was told to report only

what the CV states, to use 0 for any year that the CV does not give, and never to estimate a year, and the full prompt is printed in Appendix E.

For each CV, the model returns seven fields: the type of document, the name printed on it, a list of degrees, a list of positions held at UT Austin, whether UT Austin is mentioned at all, the earliest year of work at UT Austin, and the latest year that appears anywhere in the document. Each degree is recorded with its level, field, institution, year and status, and each position with its title, unit, role, start year and end year. The prompt gives explicit rules for the most common errors, so that a degree counts as completed only if the CV shows it as awarded, and a doctorate listed as expected or in progress is not counted. In the same way, the title Assistant Instructor is coded as a graduate student position at UT Austin, while Assistant Professor of Instruction is coded as a faculty position, and positions at other campuses of the University of Texas System are left out.

The instructor variables in Table 4 are built from this output, so that the highest degree is the highest completed degree, and the years since the degree are counted from its award year. The start year at UT Austin is the earliest start year among the positions held there, and the earliest year of work at UT Austin is used when no position gives a year. The role in each term is the role of the position held in that term, grouped into the categories shown in Table 4. The prompt was developed and checked on a random sample of 200 CVs from 2022, which were also coded by hand. The panel of CVs covers 4,441 distinct documents, which are linked to 52,624 sections from 2015 to 2026.

A.4 Fields and Subject Areas

Table A4 lists, for each of the five fields in Table 9, the course prefixes that UT Austin uses to name its subject areas. The 997 consistent courses with a 2022 syllabus come from 128 prefixes, and every prefix belongs to exactly one field.

Table A4. Course prefixes in each field, consistent courses with a 2022 syllabus

Field	Courses	Prefixes	Course prefixes
Humanities and arts	293	57	AET, AFR, AMS, ANS, ARA, ARH, ART, ASL, C C, C L, CHI, CRW, CTI, CZ, DAN, DCH, DES, E, EUS, FR, GER, GK, HEB, HIN, HIS, ITD, ITL, J S, JPN, KOR, L A, LAH, LAS, LAT, LIN, MAL, MAS, MES, MUS, NOR, PHL, POR, PRS, R S, REE, RHE, RUS, SAN, SPN, SWE, T C, T D, TUR, URD, WGS, YID, YOR
Social sciences	65	10	ANT, ECO, GOV, GRG, H S, IRG, PSY, SOC, SUS, URB
Natural sciences and mathematics	143	14	AST, BCH, BIO, C S, CH, GEO, M, MNS, NEU, NSC, P S, PHY, SDS, UTS
Engineering	125	10	ARE, ASE, BME, C E, CHE, E M, E S, EVE, M E, PGE
Professional schools	371	37	ACC, ADV, AFS, ARC, ARI, B A, CLD, CMS, COM, DEV, EDC, EDP, FIN, HDF, HDO, HED, J, KIN, LEB, M S, MAN, MIS, MKT, N, N S, NTR, O M, P R, PBH, PED, R E, R M, RTF, S W, SED, TXA, UTL
All fields	997	128	

Notes: Course prefixes are written as in the UT Austin course schedule, where some prefixes contain a space (for example, B A for business administration and C S for computer science). Courses are the consistent courses with at least one 2022 syllabus, as in Table 9.

Appendix B: Robustness Checks

The main estimates rest on several choices about when and how the treatment is measured and about which courses enter the sample, so this appendix repeats the estimation with five groups of changes to these choices, one change at a time. The first check measures both treatments from the 2019 syllabi, the next two change how the treatments are built, and the last two change the sample by keeping only in-person courses and by using the longer panel for 2015 to 2025. Some checks change the scale of the treatment, so Tables B1 and B2 report each estimate as the effect of a move from the 25th to the 75th percentile of the treatment in that sample, called an interquartile move below. Figures B1 and B2 show the event studies for the fall tracks, with the main estimate drawn next to each check, and the full tables for rows (1), (2a), (2b), (3), (4) and (5) are reported in Appendix C (Tables C1 to C6). Sections B.1 to B.5 present the five checks in turn, and Section B.6 summarizes what they show.

B.1 Base Year of the Syllabi

The main treatments are measured from the 2022 syllabi, which were written after many courses had moved part of their grade to take-home work during COVID. In this check, both treatments are measured from the 2019 syllabi of the same term, which describe how courses were graded before COVID and three years before the release of ChatGPT. These syllabi therefore cannot reflect any change that instructors made in response to COVID or to AI, and the sample keeps 486 of the 493 fall tracks (Table B1).

For AI capability, the results with the 2019 syllabi are close to the main estimates, since in the fall tracks an interquartile move in AI capability raised the share of A by 4.7 percentage points, against 5.6 percentage points in the main estimate (Table B1, Panel A). For the same move, the shares of A- and B+ fell by 1.8 and 1.1 percentage points, and mean grade points rose by 0.039 grade points. The mean within-section SD fell by 0.023 grade points, and the estimates for mean grade points and the mean within-section SD become larger and more precise when the spring tracks are added (Panel B). In the event study (Figure B1), the estimates for the share of A before 2023 lie between -2.3 and 0.8 percentage points and none of them is significant. From 2023 on, the estimates lie between 3.8 and 4.7 percentage points and are all significant at the 5 percent level.

For AI exposure, the 2019 syllabi keep the shift from A- and B+ grades to A grades but remove the rise in mean grade points. In the fall tracks, an interquartile move in AI exposure raised the share of A by 2.1 percentage points and lowered the shares of A- and B+ by 1.5 and 0.9 percentage points (Table B2, Panel A). The estimates for the share of students with A- or above, for mean grade points and for the mean within-section SD are, however, close to zero. The event study (Figure B2) also shows that the negative estimates before 2023, which appear with the 2022 syllabi, are not present with the 2019 syllabi. With the 2019 syllabi, the estimates for the share of A in these years lie between -1.0 and 0.3 percentage points. This pattern suggests that the 2022 measure of AI exposure partly picks up courses whose grading moved to take-home work, and whose grades rose, between 2020 and 2022. The main estimate for mean grade points with AI exposure should therefore be read with caution.

Table B1. Robustness checks, AI capability (2022 syllabi unless noted), effects of an interquartile move, consistent courses

	Share of A	Share of A-	Share of B+	A- or above	Mean grade points	Mean within- section SD	IQR	Courses
Panel A. Fall tracks								
Main estimate	5.6*** (1.1)	-2.2*** (0.6)	-1.1*** (0.4)	3.4*** (1.0)	0.044*** (0.017)	-0.024* (0.013)	0.663	493
(1) 2019 syllabi	4.7*** (1.1)	-1.8*** (0.5)	-1.1*** (0.4)	2.8*** (1.0)	0.039** (0.017)	-0.023* (0.013)	0.674	486
(2a) Continuous AI capability	0.5 (0.6)	0.7* (0.4)	0.0 (0.2)	1.2** (0.5)	0.018*** (0.007)	-0.010* (0.006)	0.178	493
(2b) Continuous, unclear tasks = 0	3.3*** (1.0)	-1.3** (0.5)	-0.7** (0.3)	2.0** (0.9)	0.026* (0.014)	-0.017 (0.011)	0.448	493
(2c) Problem sets AI-doable	3.1*** (1.0)	-1.3** (0.5)	-0.7** (0.3)	1.9** (0.8)	0.024* (0.014)	-0.016 (0.011)	0.521	493
(2d) Sections weighted equally	5.5*** (1.1)	-2.2*** (0.6)	-1.0*** (0.4)	3.3*** (1.0)	0.043*** (0.016)	-0.022* (0.013)	0.655	493
(4) In-person courses only	4.2*** (1.3)	-2.0*** (0.7)	-0.7 (0.5)	2.2* (1.2)	0.032 (0.020)	-0.013 (0.016)	0.661	343
(5) Balanced panel, 2015-2025	5.5*** (1.3)	-2.8*** (0.6)	-1.3*** (0.5)	2.8** (1.3)	0.027 (0.021)	-0.010 (0.014)	0.694	366
Panel B. Fall and spring tracks								
Main estimate	4.1*** (0.8)	-1.2*** (0.4)	-0.3 (0.3)	2.9*** (0.8)	0.044*** (0.012)	-0.030** (0.010)*	0.617	634
(1) 2019 syllabi	4.1*** (0.9)	-1.0** (0.4)	-0.5 (0.3)	3.1*** (0.8)	0.050*** (0.013)	-0.033** (0.011)*	0.679	625
(2a) Continuous AI capability	-0.2 (0.5)	0.6** (0.3)	0.2 (0.2)	0.4 (0.4)	0.003 (0.006)	-0.003 (0.005)	0.181	634
(2b) Continuous, unclear tasks = 0	2.2*** (0.8)	-0.6 (0.4)	0.1 (0.3)	1.6** (0.7)	0.029*** (0.011)	-0.024** (0.009)*	0.442	634
(2c) Problem sets AI-doable	2.0*** (0.8)	-0.6 (0.4)	0.1 (0.3)	1.5** (0.7)	0.027** (0.011)	-0.022** (0.009)	0.519	634
(2d) Sections weighted equally	4.1*** (0.8)	-1.3*** (0.4)	-0.3 (0.3)	2.9*** (0.8)	0.044*** (0.013)	-0.030** (0.010)*	0.619	634
(4) In-person courses only	3.6*** (1.0)	-1.0* (0.5)	-0.3 (0.4)	2.6*** (1.0)	0.036** (0.016)	-0.024* (0.013)	0.654	461
(5) Balanced panel, 2015-2025	4.4*** (1.0)	-1.7*** (0.5)	-0.7** (0.4)	2.7*** (1.0)	0.033** (0.017)	-0.022* (0.011)	0.652	464

Notes: Each estimate is the effect of a move from the 25th to the 75th percentile of the treatment, that is, the estimate of β in equation (3) multiplied by the interquartile range (IQR) of the treatment in that sample, which is given in the column IQR. Row (3) is not shown because it changes only the unproctored share, and row (2d) gives each 2022 section of a course the same weight. Shares of letter grades are in percentage points, and mean grade points and the mean within-section SD are in grade points. Each estimate comes from a separate regression with equal weights. Panel A uses the fall tracks with track and year fixed effects, and Panel B uses the fall and spring tracks with track and semester fixed effects. The main estimate uses the 2022 syllabi of the same term and the balanced panel of consistent undergraduate courses for 2018 to 2025, as in Table 10. Row (1) measures the treatment from the 2019 syllabi, row (2a) replaces the binary *AIcap_i* with the benchmark score of its task type, row (2b) does the same but sets components with no clear task to zero, and row (2c) keeps the binary rule but codes quantitative problem sets as AI-doable. Row (4) keeps the courses whose 2022 sections are all coded as in-person, and row (5) uses the balanced panel for 2015 to 2025. Standard errors, clustered by course and scaled in the same way, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B2. Robustness checks, AI exposure (2022 syllabi unless noted), effects of an interquartile move, consistent courses

	Share of A	Share of A-	Share of B+	A- or above	Mean grade points	Mean within-section SD	IQR	Courses
Panel A. Fall tracks								
Main estimate	4.5*** (0.9)	-1.8*** (0.5)	-1.1*** (0.3)	2.7*** (0.8)	0.035** (0.014)	-0.015 (0.012)	0.450	493
(1) 2019 syllabi	2.1** (0.8)	-1.5*** (0.5)	-0.9*** (0.3)	0.7 (0.7)	0.001 (0.012)	0.001 (0.011)	0.344	486
(2a) Continuous AI capability	3.3*** (0.9)	-1.2** (0.5)	-1.0*** (0.3)	2.2*** (0.8)	0.027** (0.013)	-0.007 (0.011)	0.406	493
(2b) Continuous, unclear tasks = 0	3.5*** (0.8)	-1.4*** (0.4)	-1.0*** (0.3)	2.1*** (0.7)	0.027** (0.012)	-0.013 (0.010)	0.310	493
(2c) Problem sets AI-doable	3.5*** (0.8)	-1.4*** (0.5)	-1.0*** (0.3)	2.1*** (0.7)	0.027** (0.012)	-0.013 (0.010)	0.375	493
(2d) Sections weighted equally	4.5*** (0.9)	-1.8*** (0.5)	-1.1*** (0.3)	2.7*** (0.8)	0.033** (0.014)	-0.013 (0.012)	0.450	493
(3) Indeterminate = 0.5	4.7*** (0.9)	-1.8*** (0.5)	-1.2*** (0.3)	2.9*** (0.8)	0.037** (0.014)	-0.016 (0.012)	0.471	493
(4) In-person courses only	3.4*** (1.0)	-1.5*** (0.6)	-0.8** (0.4)	1.9** (0.9)	0.023 (0.016)	-0.004 (0.014)	0.417	343
(5) Balanced panel, 2015-2025	4.9*** (1.0)	-2.1*** (0.6)	-1.5*** (0.4)	2.8*** (1.0)	0.029* (0.016)	-0.012 (0.012)	0.448	366
Panel B. Fall and spring tracks								
Main estimate	3.8*** (0.7)	-1.4*** (0.4)	-0.6** (0.2)	2.4*** (0.7)	0.036*** (0.011)	-0.024** (0.010)	0.471	634
(1) 2019 syllabi	2.1*** (0.7)	-1.3*** (0.3)	-0.5** (0.2)	0.9 (0.6)	0.009 (0.010)	-0.008 (0.009)	0.350	625
(2a) Continuous AI capability	2.6*** (0.7)	-0.9*** (0.3)	-0.5* (0.2)	1.7** (0.7)	0.026** (0.011)	-0.013 (0.009)	0.414	634
(2b) Continuous, unclear tasks = 0	2.6*** (0.6)	-0.9*** (0.3)	-0.4* (0.2)	1.7*** (0.6)	0.027*** (0.009)	-0.019** (0.008)	0.323	634
(2c) Problem sets AI-doable	2.5*** (0.6)	-0.9*** (0.3)	-0.4* (0.2)	1.7*** (0.6)	0.027*** (0.009)	-0.019** (0.008)	0.380	634
(2d) Sections weighted equally	3.8*** (0.7)	-1.4*** (0.4)	-0.6** (0.2)	2.4*** (0.7)	0.035*** (0.011)	-0.023** (0.010)	0.470	634
(3) Indeterminate = 0.5	4.0*** (0.7)	-1.4*** (0.4)	-0.6** (0.3)	2.5*** (0.7)	0.038*** (0.012)	-0.026** (0.010)	0.500	634
(4) In-person courses only	3.0*** (0.8)	-1.1** (0.5)	-0.6* (0.3)	1.9** (0.7)	0.024* (0.013)	-0.015 (0.011)	0.442	461
(5) Balanced panel, 2015-2025	4.3*** (0.8)	-1.8*** (0.5)	-1.1*** (0.3)	2.6*** (0.8)	0.030** (0.014)	-0.018* (0.010)	0.474	464

Notes: Each estimate is the effect of a move from the 25th to the 75th percentile of the treatment, that is, the estimate of β in equation (3) multiplied by the interquartile range (IQR) of the treatment in that sample, which is given in the column IQR. Shares of letter grades are in percentage points, and mean grade points and the mean within-section SD are in grade points. Each estimate comes from a separate regression with equal weights. Panel A uses the fall tracks with track and year fixed effects, and Panel B uses the fall and spring tracks with track and semester fixed effects. The main estimate uses the 2022 syllabi of the same term and the balanced panel of consistent undergraduate courses for 2018 to 2025, as in Table 10. Row (1) measures the treatment from the 2019 syllabi, row (2a) replaces the binary $Alcap_j$ with the benchmark score of its task type, row (2b) does the same but sets components with no clear task to zero, row (2c) keeps the binary rule but codes quantitative problem sets as AI-doable, and row (2d) gives each 2022 section of a course the same weight. Row (3) counts components whose proctoring is indeterminate as half unproctored. Row (4) keeps the courses whose 2022 sections are all coded as in-person, and row (5) uses the balanced panel for 2015 to 2025. Standard errors, clustered by course and scaled in the same way, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B.2 Other Ways to Build the Treatment

The binary rule codes a task type that has a benchmark as AI-doable only if GPT-4o, the free ChatGPT model in 2024-25, scored at least 70 percent on the benchmark closest to that type. In this check, the binary value of $AIcap_j$ is replaced with the score itself, so that each task type keeps the information in its benchmark result. The score is 0.83 for written prose, 0.887 for objective items, 0.902 for code or data and 0.78 for quantitative problem sets, while oral work, studio and lab work, and participation stay at zero. Components with no clear task, such as "Exam 2, 20 percent", have no benchmark, so they receive 0.873, the mean score of the three AI-doable types (row 2a). The check is then repeated with these components set to zero, as in the binary rule (row 2b).

When components with no clear task are coded as AI-doable, AI capability no longer predicts a rise in the share of A. In the fall tracks, an interquartile move in this measure raised the share of A by only 0.5 percentage points, which is not significant, and the share of A- rose rather than fell (Table B1, row 2a). The reason is that components with no clear task make up 34 percent of all graded components in the 2022 syllabi and 36 percent of the grade in the average section. Most of these components are exams that students write under a proctor, and 80 percent of the exams with no clear task are proctored when they are weighted by their share of the grade. The continuous measure therefore counts most midterms and final exams as work that AI can do, although students cannot use AI during them. As a result, almost every course receives a high value, the interquartile range falls to 0.178, and the correlation with the binary measure across tracks is only 0.08.

When components with no clear task are set to zero, the results with the benchmark scores keep the main pattern but are smaller, and they are close to those in row (2c) because this version also gives quantitative problem sets a positive score. In the fall tracks, an interquartile move in this measure raised the share of A by 3.3 percentage points and mean grade points by 0.026 grade points (Table B1, row 2b). For the same move, the shares of A- and B+ fell by 1.3 and 0.7 percentage points. The event study for this measure, which is not shown in Figure B1, has no significant estimate before 2023, and its estimates for the share of A from 2023 on lie between 3.9 and 4.9 percentage points. The failure of row (2a) therefore comes from coding proctored exams as AI-doable, not from the use of benchmark scores in place of the binary rule. Setting

quantitative problem sets to zero, while keeping 0.873 for components with no clear task, gives a much smaller rise in the share of A (1.7 percentage points, not shown in Table B1) and no fall in the share of A-.

AI exposure is much less sensitive to this choice, because it multiplies AI capability by the unproctored share and so gives proctored exams a value of zero. With components that have no clear task coded as AI-doable, an interquartile move in AI exposure raised the share of A by 3.3 percentage points in the fall tracks (Table B2, row 2a). For the same move, the shares of A- and B+ fell by 1.2 and 1.0 percentage points. The event study for this version, however, shows negative estimates for the share of A before 2023, and three of these four estimates are significant. After the release, its estimates lie between 0.6 and 1.7 percentage points and are not significant, so part of the estimate in row (2a) comes from the lower shares of A before 2023. With components that have no clear task set to zero, the estimates are similar in size (Table B2, row 2b), and no estimate before 2023 is significant.

The binary rule departs from the set of benchmarks that OpenAI reported for GPT-4o in only one place, which is the coding of quantitative problem sets. These are coded as zero on the basis of U-MATH, although GPT-4o scored 75.9 percent on MATH, the benchmark for this type in the OpenAI set (Table A1). In this check, quantitative problem sets are coded as AI-doable, as the MATH score implies, while every other task type keeps its binary code and components with no clear task stay at zero (row 2c). Problem sets carry 10.1 percent of the grade in the average fall track, and 26 percent of the fall tracks give them some weight, so the new measures stay close to the main ones, with correlations across fall tracks of 0.75 for AI capability and 0.91 for AI exposure.

With problem sets coded as AI-doable, the shift from A- and B+ grades to A grades remains for both measures, but it is smaller than in the main estimates. In the fall tracks, an interquartile move in AI exposure raised the share of A by 3.5 percentage points, against 4.5 percentage points in the main estimate (Table B2, row 2c). For the same move, the shares of A- and B+ fell by 1.4 and 1.0 percentage points and mean grade points rose by 0.027 grade points, while the effect on the mean within-section SD is small and not significant. In the fall and spring tracks, the effect of AI exposure on the share of A is 2.5 percentage points, against 3.8 percentage points in the main estimate. For AI capability, the effect on the share of A in the fall tracks falls from

5.6 to 3.1 percentage points, and the effect on mean grade points falls from 0.044 to 0.024 grade points (Table B1, row 2c).

Part of the smaller estimate for AI capability comes from the years before 2023, when the event-study estimates for this measure are positive, although none of them is significant. In this event study, which is not shown in Figure B1, the estimates for the share of A lie between 1.0 and 1.8 percentage points from 2018 to 2021, and between 3.7 and 4.7 percentage points from 2023 to 2025, when all three are significant at the 1 percent level. For AI exposure, no estimate before 2023 is significant, while the main measure has one significant estimate in 2019, and the estimates from 2023 to 2025 lie between 2.2 and 2.6 percentage points. AI exposure changes less than AI capability because it counts only the problem sets that students complete without a proctor, which make up 5.0 percent of the grade in the average fall track. This check therefore shows that the direction of the main result does not depend on how problem sets are coded, although coding them as AI-doable makes the effect on the share of A about one fifth smaller for AI exposure and almost one half smaller for AI capability.

The main treatment of each course is the mean of the exposure of its 2022 sections, weighted by the number of students in each section. In this check, each 2022 section of a course receives the same weight, so that a large section does not count more than a small one (row 2d). The two measures are almost the same, because about half of the tracks have only one section in 2022, and their correlation across fall tracks is 0.995. As a result, an interquartile move in AI exposure raised the share of A by 4.5 percentage points in the fall tracks, as in the main estimate, and the effects on the other outcomes and the event-study estimates also change very little (Table B2, row 2d). For AI capability, the share of A rose by 5.5 percentage points in this check, against 5.6 percentage points in the main estimate (Table B1, row 2d), so neither measure is sensitive to how the sections of a course are weighted.

B.3 Indeterminate Components in the Unproctored Share

The unproctored share counts a component as unproctored only when the syllabus makes it clear that students complete it without a proctor. Components whose proctoring is indeterminate, that is, components whose single weight covers both proctored and unproctored work, therefore count as proctored, and they make up 7.5 percent of all graded components and about 10 percent of the grade in the average section. In this check, each of these components counts as half

unproctored, while work done live online stays proctored. This change affects only the unproctored share, so it leaves AI capability unchanged and is reported only for AI exposure.

The new measure of AI exposure is almost the same as the main one, with a correlation of 0.995 across tracks, and the estimates are almost the same as well. In the fall tracks, an interquartile move in this measure raised the share of A by 4.7 percentage points and mean grade points by 0.037 grade points (Table B2, row 3). These estimates compare with 4.5 percentage points and 0.035 grade points in the main estimates for the same outcomes. The event study for this measure (Figure B2) follows the main event study in every year, so the treatment of indeterminate components does not change the results.

B.4 In-Person Courses

Courses that are taught online may differ from in-person courses in how their exams are given and in how their teaching changed during and after COVID. An online course may also give students more room to use AI in work that the syllabus does not describe as unproctored. In this check, the sample is limited to the tracks whose 2022 sections are all coded as in-person in the syllabi, which keeps 343 of the 493 fall tracks and 461 of the 634 courses. The sample is selected on the delivery mode in 2022 only, so it can still include courses that moved online after 2022.

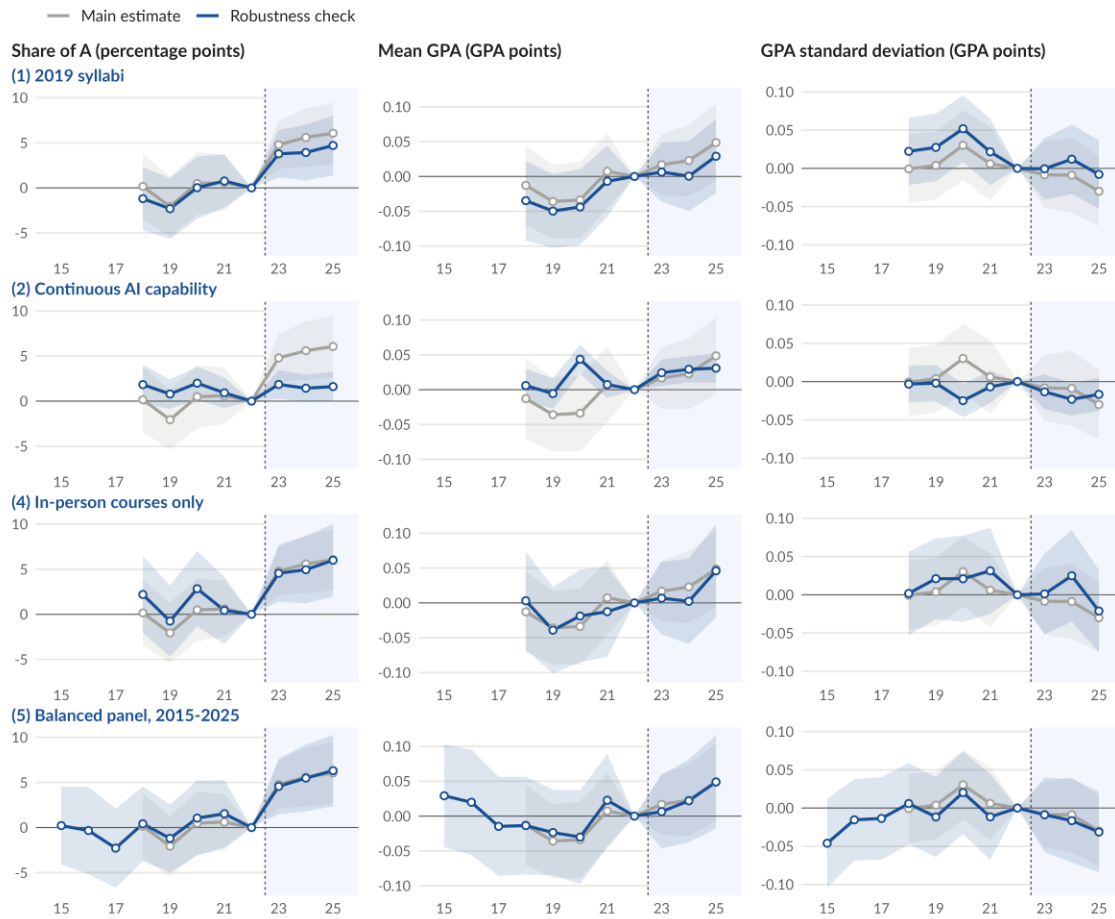
The shift from A- grades to A grades is found in the in-person courses for both measures. In the fall tracks, an interquartile move in AI capability raised the share of A by 4.2 percentage points and lowered the share of A- by 2.0 percentage points (Table B1, row 4). For AI exposure, the same move raised the share of A by 3.4 percentage points and lowered the share of A- by 1.5 percentage points (Table B2, row 4). The estimates for mean grade points and the mean within-section SD are smaller than the main estimates and are not significant in the fall tracks. For AI capability, they become significant when the spring tracks are added, with a rise in mean grade points of 0.036 grade points and a fall in the mean within-section SD of 0.024 grade points (Table B1, Panel B). The event studies (Figures B1 and B2) show no significant estimate before 2023 for either measure. The negative estimates before 2023 that appear for AI exposure in the full sample are therefore not present in the in-person courses.

B.5 Longer Panel, 2015 to 2025

The main sample covers 2018 to 2025, so it has only four years before 2022 with which to check whether courses with high and low values of the treatment followed the same path. In this check, the estimation uses the balanced panel for 2015 to 2025, which adds three earlier years but has fewer courses (464 against 634). The panel is smaller because every track must have letter grades in all eleven years, and the treatments are still measured from the 2022 syllabi.

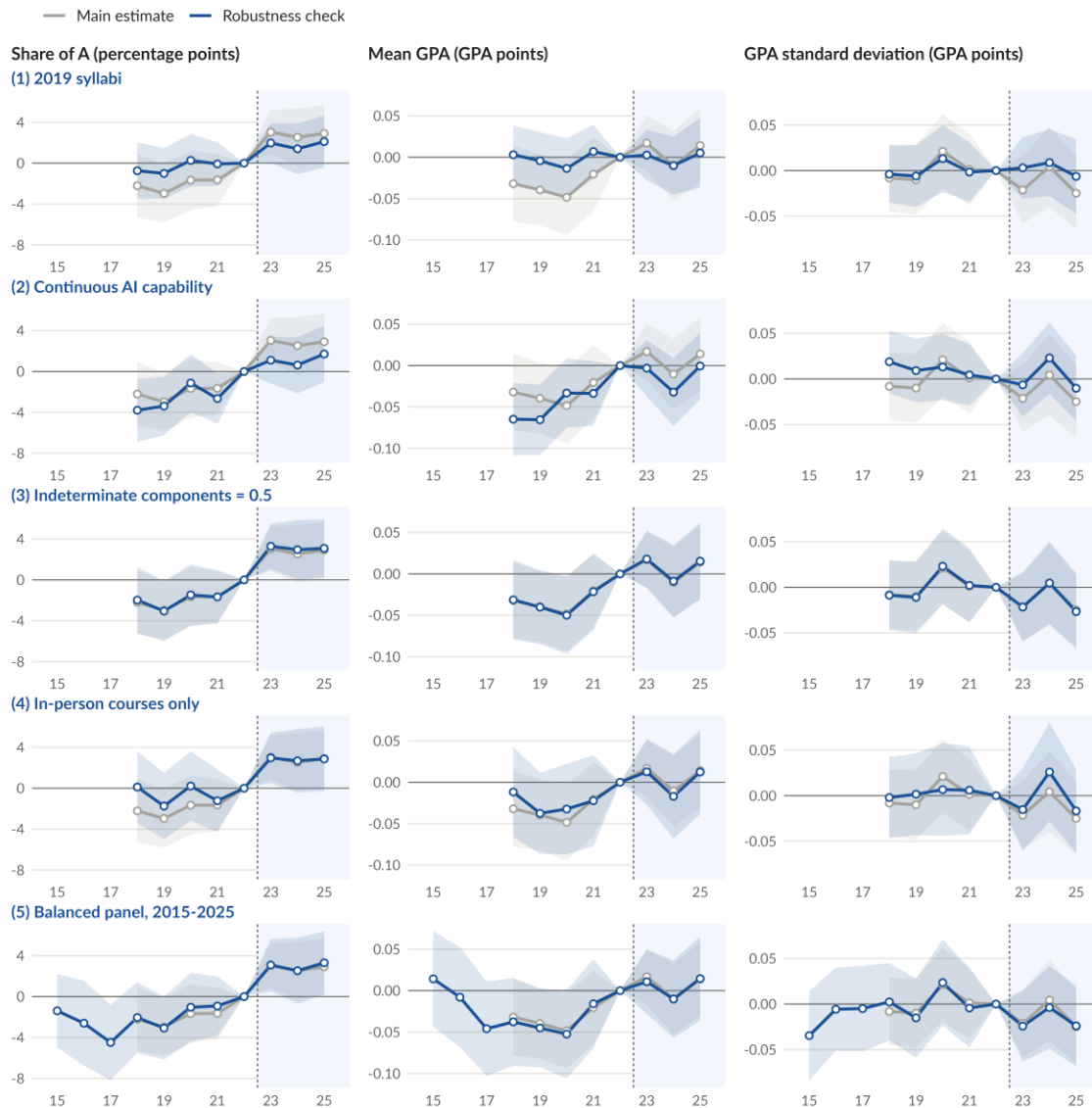
In the longer panel, the shift at the top of the distribution is about as large as in the main estimates for both measures. In the fall tracks, an interquartile move raised the share of A by 5.5 percentage points for AI capability and by 4.9 percentage points for AI exposure (Tables B1 and B2, row 5), and the shares of A- and B+ fell for both measures in both sets of tracks. For AI capability, the estimates for mean grade points and the mean within-section SD are smaller and are not significant in the fall tracks. In the fall and spring tracks, however, mean grade points rose by 0.033 grade points and the mean within-section SD fell by 0.022 grade points (Table B1, Panel B). For AI capability, the event study (Figure B1) shows no significant estimate for the share of A in any of the seven years before 2022. From 2023 on, the estimates lie between 4.6 and 6.3 percentage points and are all significant. For AI exposure, two of the seven estimates before 2022 are negative and significant (Figure B2), which repeats the pattern of the main estimates with the 2022 syllabi.

Figure B1. Robustness checks, event studies for AI capability, fall tracks of consistent courses, 2018-2025 (2015-2025 in row 5)



Notes: Each point is the estimate of β_k in equation (4) for the outcome named at the top of the column, relative to 2022, multiplied by the interquartile range of the treatment in that sample. Each point therefore shows the effect of a move from the 25th to the 75th percentile of AI capability. The grey line repeats the main estimate (2022 syllabi, balanced panel of consistent courses, 2018-2025), and the navy line shows the check named at the left of the row. Row (1) measures AI capability from the 2019 syllabi, and row (2a) uses the benchmark score of each task type, with 0.873 for components that have no clear task. Row (4) keeps the courses whose 2022 sections are all in-person, and row (5) uses the balanced panel for 2015 to 2025. The share of A is in percentage points, and mean grade points and the mean within-section SD are in grade points. Panels in the same column share one y scale. All estimates use the fall tracks, track and year fixed effects and equal weights. The shaded area marks the years after the release of ChatGPT in November 2022, and shaded bands are 95 percent confidence intervals based on standard errors clustered by course.

Figure B2. Robustness checks, event studies for AI exposure, fall tracks of consistent courses, 2018-2025 (2015-2025 in row 5)



Notes: As in Figure B1, for AI exposure. Row (2a) uses the benchmark score of each task type in AI exposure, and row (3) counts components whose proctoring is indeterminate as half unproctored.

B.6 Summary of the Robustness Checks

Across the five checks, the shift from A- and B+ grades to A grades is the most robust result of this paper. It appears for AI exposure in every check, and for AI capability in every check except the version that counts proctored exams with no clear task as work that AI can do. Coding quantitative problem sets as AI-doable makes this shift smaller, by about one fifth for AI exposure and almost one half for AI capability, but it remains significant for both measures, while giving each 2022 section the same weight leaves the results for AI exposure almost

unchanged. The rise in mean grade points and the fall in the mean within-section SD point in the same direction in almost every check, but they are smaller and less precise in the in-person courses and in the longer panel. For AI exposure, they disappear when the treatment is measured from the 2019 syllabi. The negative estimates before 2023 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, where the effects on the share of A are smaller but still significant. Taken together, the checks support the view that AI raised grades mainly by moving students from just below the top grade to the top grade. They also show that this effect is linked to graded work that AI can do and that students complete without a proctor.

Appendix C: Additional Robustness Results

This appendix reports the full results of the robustness checks in Appendix B, in the same layout as Table 10, with the coefficients for a one-unit change in the treatment and all outcomes.

Table C1. Difference-in-differences estimates, AI capability and AI exposure (2019 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
<i>Panel A. Shares of letter grades</i>				
Share of A	0.069*** (0.016) [0.444]	0.060*** (0.013) [0.462]	0.062** (0.024) [0.444]	0.061*** (0.019) [0.462]
Share of A-	-0.027*** (0.008) [0.163]	-0.015** (0.006) [0.158]	-0.043*** (0.013) [0.163]	-0.036*** (0.010) [0.158]
Share of B+	-0.017*** (0.006) [0.105]	-0.007 (0.004) [0.103]	-0.027*** (0.009) [0.105]	-0.015** (0.007) [0.103]
<i>Panel B. Cumulative shares</i>				
A- or above	0.042*** (0.015) [0.606]	0.046*** (0.012) [0.621]	0.019 (0.020) [0.606]	0.025 (0.017) [0.621]
B+ or above	0.025* (0.014) [0.711]	0.039*** (0.011) [0.723]	-0.008 (0.018) [0.711]	0.009 (0.015) [0.723]
B or above	0.015 (0.010) [0.828]	0.028*** (0.008) [0.838]	-0.015 (0.014) [0.828]	0.002 (0.011) [0.838]
<i>Panel C. Mean and dispersion of grade points</i>				
Mean grade points	0.058** (0.025) [3.409]	0.073*** (0.020) [3.433]	0.002 (0.035) [3.409]	0.024 (0.028) [3.433]
Mean within-section SD	-0.035* (0.019) [0.664]	-0.049*** (0.015) [0.650]	0.004 (0.031) [0.664]	-0.024 (0.025) [0.650]
<i>Panel D. Low grades</i>				
Share of F	-0.002 (0.002) [0.015]	0.000 (0.002) [0.016]	-0.001 (0.003) [0.015]	0.002 (0.003) [0.016]
Share of D or F	-0.004 (0.003) [0.032]	-0.005* (0.003) [0.031]	0.002 (0.005) [0.032]	0.000 (0.004) [0.031]
Course-semesters	3,888	7,184	3,888	7,184
Courses	486	625	486	625
Letter-grade observations	688,807	1,240,828	688,807	1,240,828

Notes: As in Table 10, with the treatment measured from the 2019 syllabi of the same term. The sample is limited to tracks with a coded 2019 syllabus. The interquartile range of the treatment is 0.674, 0.679, 0.344 and 0.350 in columns (1) to (4). * p<0.10, ** p<0.05, *** p<0.01.

Table C2. Difference-in-differences estimates, continuous 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	0.030 (0.034) [0.446]	−0.013 (0.030) [0.463]	0.082*** (0.022) [0.446]	0.063*** (0.017) [0.463]
Share of A-	0.037* (0.022) [0.162]	0.033** (0.015) [0.158]	−0.028** (0.012) [0.162]	−0.022*** (0.008) [0.158]
Share of B+	0.001 (0.010) [0.104]	0.010 (0.010) [0.102]	−0.024*** (0.007) [0.104]	−0.011* (0.006) [0.102]
Panel B. Cumulative shares				
A- or above	0.067** (0.026) [0.608]	0.020 (0.024) [0.622]	0.054*** (0.020) [0.608]	0.041** (0.016) [0.622]
B+ or above	0.068*** (0.022) [0.712]	0.030 (0.018) [0.724]	0.030 (0.019) [0.712]	0.030** (0.015) [0.724]
B or above	0.030* (0.016) [0.829]	−0.008 (0.013) [0.838]	0.022* (0.012) [0.829]	0.024** (0.010) [0.838]
Panel C. Mean and dispersion of grade points				
Mean grade points	0.101*** (0.038) [3.411]	0.014 (0.034) [3.434]	0.068** (0.033) [3.411]	0.062** (0.026) [3.434]
Mean within-section SD	−0.058* (0.035) [0.664]	−0.014 (0.029) [0.649]	−0.017 (0.027) [0.664]	−0.031 (0.021) [0.649]
Panel D. Low grades				
Share of F	−0.007** (0.003) [0.015]	−0.004* (0.002) [0.016]	0.001 (0.003) [0.015]	0.001 (0.002) [0.016]
Share of D or F	−0.007 (0.005) [0.032]	0.002 (0.004) [0.031]	−0.000 (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: As in Table 10, with $AIcap_j$ equal to the benchmark score of its task type for GPT-4o in 2024-25: 0.83 for written prose, 0.887 for objective items, 0.902 for code or data and 0.78 for quantitative problem sets, with 0.873, the mean of the three AI-doable types, for components with no clear task, and zero for oral work, studio and lab work, and participation. Columns (3) and (4) use this $AIcap_j$ in AI exposure. The interquartile range of the treatment is 0.178, 0.181, 0.406 and 0.414 in columns (1) to (4). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C3. Difference-in-differences estimates, continuous AI capability with unclear tasks set to zero (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
<i>Panel A. Shares of letter grades</i>				
Share of A	0.074*** (0.022) [0.446]	0.049*** (0.017) [0.463]	0.111*** (0.025) [0.446]	0.080*** (0.019) [0.463]
Share of A-	-0.029** (0.012) [0.162]	-0.013 (0.008) [0.158]	-0.044*** (0.014) [0.162]	-0.027*** (0.010) [0.158]
Share of B+	-0.016** (0.008) [0.104]	0.002 (0.006) [0.102]	-0.033*** (0.009) [0.104]	-0.012* (0.006) [0.102]
<i>Panel B. Cumulative shares</i>				
A- or above	0.045** (0.019) [0.608]	0.036** (0.016) [0.622]	0.067*** (0.022) [0.608]	0.053*** (0.017) [0.622]
B+ or above	0.029 (0.018) [0.712]	0.037*** (0.014) [0.724]	0.034 (0.021) [0.712]	0.041*** (0.016) [0.724]
B or above	0.016 (0.012) [0.829]	0.025** (0.010) [0.838]	0.025* (0.015) [0.829]	0.034*** (0.011) [0.838]
<i>Panel C. Mean and dispersion of grade points</i>				
Mean grade points	0.058* (0.031) [3.411]	0.066*** (0.025) [3.434]	0.088** (0.038) [3.411]	0.085*** (0.029) [3.434]
Mean within-section SD	-0.038 (0.025) [0.664]	-0.053*** (0.021) [0.649]	-0.043 (0.033) [0.664]	-0.060** (0.025) [0.649]
<i>Panel D. Low grades</i>				
Share of F	-0.003 (0.003) [0.015]	-0.003 (0.002) [0.016]	-0.002 (0.003) [0.015]	-0.000 (0.002) [0.016]
Share of D or F	-0.002 (0.004) [0.032]	-0.005 (0.003) [0.031]	-0.002 (0.006) [0.032]	-0.004 (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: As in Table C2, with components that have no clear task set to zero. The interquartile range of the treatment is 0.448, 0.442, 0.310 and 0.323 in columns (1) to (4). * p<0.10, ** p<0.05, *** p<0.01.

Table C4. Difference-in-differences estimates, AI exposure with indeterminate components counted as half unproctored (2022 syllabi), balanced panel of consistent courses, 2018-2025, unweighted

Treatment Semesters	(1) AI exposure Fall	(2) AI exposure Fall and spring
<i>Panel A. Shares of letter grades</i>		
Share of A	0.101*** (0.020) [0.446]	0.079*** (0.015) [0.463]
Share of A-	-0.039*** (0.011) [0.162]	-0.028*** (0.008) [0.158]
Share of B+	-0.025*** (0.007) [0.104]	-0.012** (0.005) [0.102]
<i>Panel B. Cumulative shares</i>		
A- or above	0.062*** (0.018) [0.608]	0.051*** (0.014) [0.622]
B+ or above	0.037** (0.017) [0.712]	0.039*** (0.013) [0.724]
B or above	0.023* (0.012) [0.829]	0.028*** (0.009) [0.838]
<i>Panel C. Mean and dispersion of grade points</i>		
Mean grade points	0.078** (0.031) [3.411]	0.077*** (0.023) [3.434]
Mean within-section SD	-0.033 (0.026) [0.664]	-0.052** (0.020) [0.649]
<i>Panel D. Low grades</i>		
Share of F	0.000 (0.003) [0.015]	0.001 (0.002) [0.016]
Share of D or F	-0.001 (0.005) [0.032]	-0.003 (0.004) [0.031]
Course-semester	3,944	7,272
Courses	493	634
Letter-grade observations	693,259	1,249,177

Notes: As in columns (3) and (4) of Table 10, with *Unproc_i* equal to 0.5 for components whose proctoring is indeterminate. Work done live online stays proctored. The interquartile range of the treatment is 0.471 and 0.500 in columns (1) and (2). * p<0.10, ** p<0.05, *** p<0.01.

Table C5. Difference-in-differences estimates, AI capability and AI exposure (2022 syllabi), in-person courses, 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
<i>Panel A. Shares of letter grades</i>				
Share of A	0.064*** (0.020) [0.437]	0.055*** (0.016) [0.455]	0.081*** (0.023) [0.437]	0.067*** (0.018) [0.455]
Share of A-	-0.030*** (0.011) [0.160]	-0.015* (0.008) [0.156]	-0.036*** (0.013) [0.160]	-0.025** (0.010) [0.156]
Share of B+	-0.011 (0.007) [0.108]	-0.004 (0.006) [0.104]	-0.019** (0.009) [0.108]	-0.013* (0.007) [0.104]
<i>Panel B. Cumulative shares</i>				
A- or above	0.034* (0.018) [0.598]	0.039*** (0.015) [0.611]	0.045** (0.021) [0.598]	0.043** (0.017) [0.611]
B+ or above	0.023 (0.016) [0.705]	0.035*** (0.013) [0.716]	0.026 (0.020) [0.705]	0.030** (0.015) [0.716]
B or above	0.012 (0.012) [0.826]	0.023** (0.010) [0.835]	0.015 (0.015) [0.826]	0.020* (0.011) [0.835]
<i>Panel C. Mean and dispersion of grade points</i>				
Mean grade points	0.048 (0.030) [3.401]	0.056** (0.024) [3.422]	0.056 (0.037) [3.401]	0.053* (0.029) [3.422]
Mean within-section SD	-0.020 (0.024) [0.662]	-0.037* (0.020) [0.649]	-0.011 (0.033) [0.662]	-0.035 (0.026) [0.649]
<i>Panel D. Low grades</i>				
Share of F	-0.001 (0.003) [0.015]	0.002 (0.002) [0.016]	0.001 (0.004) [0.015]	0.003 (0.003) [0.016]
Share of D or F	-0.003 (0.004) [0.032]	-0.001 (0.004) [0.031]	-0.002 (0.005) [0.032]	-0.001 (0.005) [0.031]
Course-semesters	2,744	4,680	2,744	4,680
Courses	343	461	343	461
Letter-grade observations	354,102	590,142	354,102	590,142

Notes: As in Table 10, for the tracks whose 2022 sections are all coded as in-person in the syllabi. The interquartile range of the treatment is 0.661, 0.654, 0.417 and 0.442 in columns (1) to (4). * p<0.10, ** p<0.05, *** p<0.01.

Table C6. Difference-in-differences estimates, AI capability and AI exposure (2022 syllabi), balanced panel of consistent courses, 2015-2025, unweighted

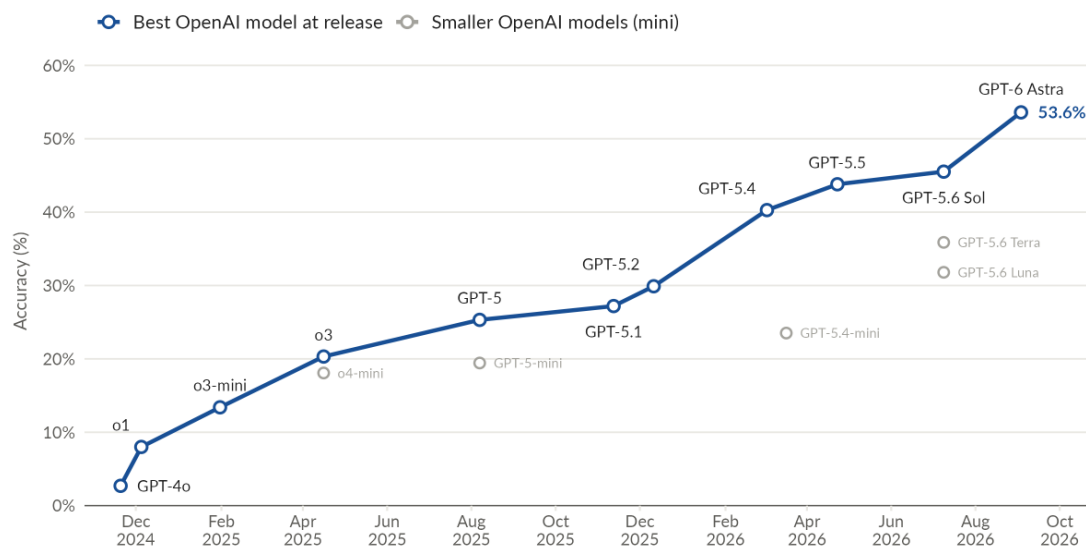
Treatment Semesters	(1) AI capability Fall	(2) AI capability Fall and spring	(3) AI exposure Fall	(4) AI exposure Fall and spring
<i>Panel A. Shares of letter grades</i>				
Share of A	0.080*** (0.019) [0.393]	0.067*** (0.015) [0.405]	0.110*** (0.022) [0.393]	0.092*** (0.017) [0.405]
Share of A-	-0.040*** (0.009) [0.163]	-0.026*** (0.007) [0.159]	-0.048*** (0.013) [0.163]	-0.037*** (0.010) [0.159]
Share of B+	-0.019*** (0.007) [0.114]	-0.011** (0.005) [0.112]	-0.032*** (0.008) [0.114]	-0.023*** (0.006) [0.112]
<i>Panel B. Cumulative shares</i>				
A- or above	0.040** (0.018) [0.557]	0.041*** (0.015) [0.564]	0.062*** (0.021) [0.557]	0.054*** (0.017) [0.564]
B+ or above	0.021 (0.017) [0.670]	0.030** (0.014) [0.676]	0.030 (0.021) [0.670]	0.032* (0.017) [0.676]
B or above	0.009 (0.013) [0.804]	0.019* (0.011) [0.809]	0.014 (0.016) [0.804]	0.019 (0.012) [0.809]
<i>Panel C. Mean and dispersion of grade points</i>				
Mean grade points	0.039 (0.030) [3.339]	0.051** (0.026) [3.350]	0.065* (0.036) [3.339]	0.063** (0.029) [3.350]
Mean within-section SD	-0.014 (0.020) [0.687]	-0.034* (0.018) [0.681]	-0.028 (0.026) [0.687]	-0.037* (0.021) [0.681]
<i>Panel D. Low grades</i>				
Share of F	0.003 (0.002) [0.016]	0.003 (0.002) [0.017]	0.002 (0.003) [0.016]	0.003 (0.002) [0.017]
Share of D or F	0.003 (0.004) [0.036]	0.001 (0.003) [0.036]	0.002 (0.005) [0.036]	0.001 (0.004) [0.036]
Course-semesters	4,026	7,447	4,026	7,447
Courses	366	464	366	464
Letter-grade observations	816,588	1,465,147	816,588	1,465,147

Notes: As in Table 10, for the balanced panel of consistent courses for 2015 to 2025, in which every track has letter grades in all eleven years from 2015 to 2025. The interquartile range of the treatment is 0.694, 0.652, 0.448 and 0.474 in columns (1) to (4). * p<0.10, ** p<0.05, *** p<0.01.

Appendix D: Additional Figure

Figure D1 shows the accuracy of OpenAI models on Humanity's Last Exam from November 2024 to September 2026, which footnote 2 of the paper uses to compare GPT-4o with the models that students can use today.

Figure D1. Accuracy of OpenAI models on Humanity's Last Exam, November 2024 to September 2026



Notes: Humanity's Last Exam has 2,500 closed-ended questions in more than 100 subjects, written by about 1,000 experts. The navy line connects the best OpenAI model at each release date, and grey points mark the smaller (mini) models; nano models are not shown. Scores are from the AI Dashboard at lastexam.ai (Center for AI Safety and Scale AI 2026), accessed on September 30, 2026.

Appendix E: Prompts

This appendix reproduces the two prompts used in this paper. The syllabus prompt was sent to Gemini 3.7 Flash for the course panel and to GLM-5.3 for the instructor panel, and the CV prompt was sent to Qwen3.6-27B. Each prompt is printed as it was sent, except that the names of the instructors in the worked examples of the CV prompt are replaced by placeholders such as [Name A].

Table E1 lists the settings of the three coding runs in one place, so that each run can be repeated with the prompts printed below.

Table E1. Model settings for the coding of syllabi and instructor CVs

Setting	Syllabi, course panel	Syllabi, instructor panel	Instructor CVs
Model	Gemini 3.7 Flash	GLM-5.3 (thinking, high)	Qwen3.6-27B
Access	Google batch services, through OpenRouter and Google AI Studio	Language-model service of Virginia Tech ARC	vLLM on one NVIDIA A100 GPU at Virginia Tech ARC
Reasoning level	Medium	High	Off
Temperature	0	0	0
Seed	20260812	20260812	42
Maximum output	24,000 tokens	24,000 tokens (streamed)	3,072 tokens
Input	Text of one syllabus; scanned syllabi as page images	Text of one syllabus; 79 scanned syllabi left out	Text of one CV, up to a context window of 49,152 tokens
Output format	Fixed JSON schema (Table A3)	Same JSON schema	JSON object with seven fields
Prompt	Appendix E.1	Appendix E.1	Appendix E.2
Run dates	August and September 2026	September 2026	August 2026
Units coded	About 11,800 (2,590 for 2022, 2,110 for 2019, 7,065 for other years)	12,264 of 12,265	4,441 CV documents

Notes: A unit is one coding request, after identical documents shared by several sections were combined. The maximum output includes the reasoning tokens of the model. A temperature of 0 and a fixed seed make repeated requests return the same answer as far as each service allows. ARC is the Advanced Research Computing center of Virginia Tech.

E.1 Syllabus Coding Prompt

The system message of each request contains the preamble followed by the rulebook, and the user message contains the text of one syllabus, preceded by the line "REGISTERED COURSE NUMBER: <course number>". The answer must follow a fixed JSON schema whose fields are listed in Table A3. Both texts were first written for a comparison with an earlier coder, which is why they refer to another coder, and they were kept unchanged in every production run so that all runs used identical instructions.

Preamble

You are coding a single university course syllabus against a fixed rulebook. Another system has already coded this document; you will not see its answers. Code the document from its text alone.

Return ONE JSON object with exactly the seven fields specified below: `course_level`, `delivery_mode`, `online_modality`, `class_sessions`, `assignments`, `ai_policy`, `grading_basis`. Do not return `unproctored_share` -- it is computed downstream from your `assignments` by the same code that scores every other coder, so that all sides are scored identically.

Each assignment also carries `task_type`, `task_evidence` and `evidence_location`. `task_type` says WHAT the work is; `administration` and `type` say WHERE it happens and whether anyone watches. They are separate questions and must not influence each other. When the syllabus does not describe what a component consists of, "unspecified" is the correct answer -- do not guess from the department.

Rulebook

The rules (identical to the ones the first coder was given; worked examples withheld)

First, check whether this document is actually a course syllabus. If it appears to be something else (a CV, resume, reading list, course schedule, or any other non-syllabus document), return exactly:

```
{ "course_level": "not_a_syllabus", "delivery_mode": "unknown", "online_modality": "unknown",  
  "class_sessions": { "in_person_sessions": false, "online_sessions": false,  
    "hyflex_student_choice": false, "temporary_remote_only": false }, "assignments": [], "ai_policy":  
  "no", "grading_basis": "unknown" }
```

NOTES ON THE INPUT:

- * A line "REGISTERED COURSE NUMBER: <number>" may appear before the syllabus text. If present, this is the specific course being analyzed – use it to resolve any ambiguity about course level or which grading scheme applies.
- * Very long syllabi may have their middle pages omitted, marked by "[... MIDDLE PAGES OF SYLLABUS OMITTED ...]". Base your answers on the visible text; do not guess about omitted content.

Otherwise, analyze this syllabus and return a JSON object with exactly these seven fields (`course_level`, `delivery_mode`, `online_modality`, `class_sessions`, `assignments`, `ai_policy`, `grading_basis`):

1. `"course_level"`: "graduate", "undergraduate", or "unknown"
 - * If a REGISTERED COURSE NUMBER is provided: 100-499 = undergraduate, 500+ = graduate.
 - * Otherwise: undergraduate if the syllabus course number is 100-499 or it says undergraduate/lower-division/upper-division; graduate if 500+ or it says graduate/master/doctoral/PhD.
 - * unknown: cannot determine
2. `"delivery_mode"`: "in-person", "online", "hybrid", or "unknown"
 - * Base this ONLY on how scheduled class/lecture sessions are conducted – ignore online exams, homework, assignments, Canvas submissions, or any other non-lecture course components.
 - * in-person: all scheduled class sessions meet physically on campus. Online components such as exams, assignments, or Canvas submissions do not change this classification.
 - * online: all scheduled class sessions are conducted online. No physical attendance required.
 - * hybrid: some scheduled class sessions are in-person AND some are online.
 - * unknown: cannot determine.
 - * HYFLEX: if students may attend the SAME scheduled sessions EITHER in person OR online at their own choice, classify as "hybrid".
 - * SELF-LABELED "HYBRID": a syllabus that calls itself "hybrid" is only hybrid if its scheduled CLASS SESSIONS are split between in-person and online. If the "hybrid"/"online" part is just independent assignment hours, Canvas work, or online homework while ALL class meetings are in person, it is "in-person".
 - * DOES NOT COUNT AS ONLINE SESSIONS: posted recordings of in-person lectures, a standby/backup Zoom link for an otherwise in-person class, or online homework/exams/Canvas submissions. These do NOT make a course hybrid or online.
 - * TEMPORARY EMERGENCY-REMOTE: if class sessions are online ONLY for a temporary, stated emergency period (e.g., "the first two weeks will be on Zoom due to COVID/Omicron, then we return to in person"), the course is "in-person" – the temporary displacement does not make it hybrid.
 - * CRITICAL – MULTI-SECTION SYLLABI: Some syllabi cover multiple sections with different delivery modes (e.g., "Section 001: Face-to-face in UTC 4.102" and "Section 002: Internet (asynchronous)"). Do NOT classify such a course as "hybrid". Each section has its own delivery mode. If a REGISTERED COURSE NUMBER is provided and the syllabus identifies which section it corresponds to, use that section's delivery mode. Otherwise, if you cannot

determine a single delivery mode that applies to all sections, return "unknown". HOWEVER, if the sections differ only in ONLINE MODALITY – for example one section synchronous online and another asynchronous online, with no section meeting physically on campus – then a single delivery mode DOES apply and you must return it: "online". Reserve "unknown" for genuine delivery-mode conflicts, such as one section face-to-face and another internet-only.

* 2b. "class_sessions" – ALSO report these four OBSERVABLE booleans about the scheduled class meetings (delivery_mode is cross-checked against them):

- "in_person_sessions": true if ANY scheduled class meeting is held physically on campus.
- "online_sessions": true if ANY scheduled class delivery happens online in a way that REPLACES an in-person meeting – live online class sessions, SPECIFIC class sessions listed as held on Zoom (e.g., "Jan 18 and 20 via Zoom", "at least three classes on Zoom"), a course/section DESIGNATED "Internet"/"online"/"web-based", OR asynchronous online modules/recorded lectures that ARE the class. FALSE only for mere supplementary recordings of an in-person lecture, a standby/backup Zoom link for an otherwise in-person class, or online homework/exams.

- "hyflex_student_choice": true if students may attend the same session in person OR online at their own choice (e.g., "come to the room OR join by Zoom"), INCLUDING when that option persists for the rest of the term after a remote start.

- "temporary_remote_only": true ONLY if online class sessions are strictly time-boxed to a short, explicitly stated emergency period (e.g., "the first two weeks") AFTER WHICH ALL sessions are held in person with NO ongoing online option. This is FALSE if: (a) any online attendance option continues for the rest of the term (set hyflex_student_choice or online_sessions instead), (b) the online period is open-ended or whole-semester (e.g., "on Zoom until vaccinations", "online until further notice", "all classes on Zoom" – set online_sessions true and in_person_sessions false), or (c) only scattered individual sessions are on Zoom rather than a single contiguous opening block (set online_sessions true).

3. "online_modality": "asynchronous", "synchronous", "mixed", "n/a", or "unknown"

* Describes how the ONLINE portion of the course is delivered. Only meaningful when delivery_mode is "online" or "hybrid".

* "asynchronous": no scheduled live meetings. Content is recorded lectures, weekly modules, or self-paced work. Signals: "asynchronous", "self-paced", "no scheduled meeting times", "work at your own pace", "recorded lectures", "modules due by [date]".

* "synchronous": scheduled live class sessions students must attend in real time (via Zoom, Teams, Webex, etc.). Signals: "synchronous", "live class", "Zoom meetings at [day/time]", "scheduled class meetings", "live sessions required", "attend in real time".

* "mixed": the online portion has BOTH asynchronous and synchronous components (e.g., recorded lectures plus a required weekly live discussion).

* "n/a": delivery_mode is "in-person" – there is no online component to classify.

* "unknown": delivery_mode is "online" or "hybrid" but the syllabus does not state whether sessions are live or self-paced. Also use "unknown" when delivery_mode itself is "unknown".

4. "assignments": a list of objects detailing the grading breakdown. STEP 1: Find the grading or grade breakdown section. If there is NO breakdown, return an empty list: [] SPECIAL CASE – COMPLETION-BASED PASS/FAIL: if the course is pass/fail and passing simply requires completing all required components (no weights or points given), list each required component with "weight": 1 and "weight_unit": "count". STEP 2: List every graded component. **BONUS AND EXTRA CREDIT ARE NOT GRADED COMPONENTS.** Exclude any row the syllabus describes as bonus points, extra credit, or points that can only be added on top of the grade, EVEN IF it appears inside the grading table. The components you list should be the ones that make up the grade (normally summing to 100% or to the stated point total); anything that only adds on top is excluded. STEP 3: For each component record its "name", its "weight" (as a single resolved integer or decimal number – never a formula or expression) and its "weight_unit" ("percent", "points", or "count"). Do not attempt to calculate percentages if only points are given.

* Weight must always be a plain number (e.g., 25, not "25% x 2", not "10% x 1").

* If the syllabus expresses a weight as a multiplication (e.g., "Exams (10% x 1, 25% x 2)"), expand into separate entries with individual weights: one entry at 10 and two entries at 25.

* If the syllabus says "N assignments at X% each" (e.g., "Case Study Memoranda (4): 12.5% each"), list N separate entries each with weight X – do NOT sum them into one entry.

* If the same grade component appears in both a percentage table AND a points table (dual representation), use only one representation – prefer percentages if both are given. Do NOT list the same assignment twice under different units.

STEP 4 – for each component record TWO fields, "administration" then "type".

>>> 4a. "administration" – report the OBSERVABLE FACT of WHERE / HOW the component is completed. This is not a judgment about AI; it is simply where it happens. Choose one:

* "in_person": completed in a supervised physical setting on campus – in-class exam/quiz/midterm/final, in-class participation, attendance, lab practical, field exam.

* "online_live": completed online during a scheduled LIVE session the instructor observes in real time – live synchronous online participation, or an exam taken live in a proctored Zoom/Teams session.

* "online_proctored": completed online AND the syllabus NAMES a webcam/lockdown proctoring tool for THIS SPECIFIC component (Proctorio, Honorlock, Respondus LockDown/Monitor, ProctorU,

Examity, ExamSoft) or requires "camera on"/"webcam required" for it. Attribute PER COMPONENT: if a tool is named for the final only, the final is "online_proctored" but the tool-free quizzes are "online_self". Do NOT blanket-apply a tool named for one exam to the others.

* "online_self": completed online on the student's own, with NO named proctoring tool – a Canvas/online quiz or exam taken within a time window, an online discussion post. (Being timed or closed-book still counts as online_self – a time limit is NOT observation.) **A LEARNING-PLATFORM NAME IS NOT A LOCATION.** Canvas, Blackboard, Gradescope, Quest, Connect and the like say where work is delivered or submitted, not where the student sits. If the syllabus ties the component to class time, a scheduled face-to-face session, or a named room, it is "in_person" even though it runs through the platform ("administered during class time and on Canvas" = in_person; "in-class Canvas quizzes" = in_person).

* "take_home": a take-home assignment, paper, essay, project, portfolio, problem set, or take-home exam. **SUBMISSION LANGUAGE vs SITTING LANGUAGE, for anything called an exam, test or quiz.** An exam is "take_home" when the syllabus hands out its prompt, guidelines or topic in advance AND the exam is "due" at a clock time, submitted, or uploaded -- this holds even in a fully in-person course, and even when the item is called an exam. An exam is "in_person" when the syllabus locates it in class time, in a named room, or in a scheduled or reserved on-campus slot (including lab practicals and partnered skills demonstrations). Apply the delivery-conditional default below ONLY when the syllabus gives neither signal.

* "n/a": administration is not a meaningful single event – e.g., a single combined bucket that mixes several kinds of work (these will be "indeterminate" below).

* "unknown": the syllabus does not make the administration clear.

DELIVERY-CONDITIONAL DEFAULT (apply the intended course delivery unless the syllabus says otherwise for THAT specific component):

- If the course is in-person (in_person_sessions true AND (online_sessions false OR temporary_remote_only true)): assume exams/quizzes/tests are "in_person" unless the syllabus explicitly says a specific one is online or take-home. A TEMPORARY emergency-remote opening block does not change this: if the first two or three class meetings are held on Zoom and the course then returns to the classroom for the rest of the term, exams remain "in_person" unless that specific exam is itself stated to be online or take-home.

- If the course has ANY ONGOING online sessions ((online_sessions true AND temporary_remote_only false) OR hyflex_student_choice true – i.e., genuinely online or hybrid, NOT a course that merely opened with a short emergency-remote block): assume exams/quizzes/tests are "online_self" unless the syllabus, for THAT specific component, names a webcam/lockdown tool (then "online_proctored"), explicitly says that exam is taken on campus / in class (then "in_person"), or is a live proctored session (then "online_live").

- CRITICAL: do NOT set an exam's administration to "in_person" merely because lectures meet on campus part of the time. Mark an exam "in_person" ONLY when the syllabus actually locates that exam in a supervised on-campus/in-class setting. If a course has ONGOING online sessions (temporary_remote_only false) and the exam's location is unstated, it is "online_self", NOT "in_person". But if the course's ONLY online sessions were a temporary emergency-remote block (temporary_remote_only true) and the exam's location is unstated, the default is "in_person".

>>> 4b. "type" – "proctored", "unproctored", or "indeterminate". Determine it from the administration you just recorded, using this GATE (it OVERRIDES any intuition that "exams are proctored"):

* "in_person", "online_live", or "online_proctored" -> "proctored" (supervised in real time, or a named webcam/lockdown tool; AI cannot realistically be used).

* "online_self" -> "unproctored" – EVEN IF the exam is timed, closed-book, or a "final". A time limit is NOT supervision. (If a webcam/lockdown tool IS named for it, you should have set administration to "online_proctored" above, not "online_self".)

* "take_home" -> "unproctored".

* "indeterminate": use ONLY when a SINGLE combined weight covers a MIX of proctored AND unproctored sub-components with no way to split the weight (set administration to "n/a"). Do NOT use indeterminate merely because you are unsure. **PREPARATION IS NOT A SUB-COMPONENT.** Almost every supervised assessment is prepared for beforehand; that does NOT make it a mix. Use indeterminate only when the SYLLABUS ITSELF describes the component as containing both supervised and unsupervised GRADED work (e.g. "in addition to an in-class portion, exams may also include some take-home problems", or "participation includes both in person and remote training", or "pop quizzes in class or take-home quizzes done outside of class"). A component that is simply supervised -- attendance, a clinical or field placement, an in-class practical, a lab or skills or performance exam held in a scheduled or reserved on-campus slot -- is "in_person", NOT indeterminate, even though students prepare for it outside class. Likewise a paper, essay, plan or study submitted for a grade is "take_home", NOT indeterminate.

NOTES:

* Attendance, in-class discussion, and in-class participation are "in_person" -> PROCTORED.

* Presentations are INDETERMINATE – delivery is supervised but the content/preparation is fully AI-boostable (administration "n/a"). **This applies ONLY to components the syllabus itself calls a presentation, talk, performance, recital or critique.** Do NOT extend this preparation

reasoning to any other supervised component; see the PREPARATION IS NOT A SUB-COMPONENT rule above.

- * Lab exams and field exams are supervised assessments -> "in_person" -> PROCTORED. A standalone take-home "lab report" is "take_home" -> UNPROCTORED. A "lab exam report" documenting a supervised practical is INDETERMINATE. Field projects/field trips are INDETERMINATE.

- * Quizzes and tests follow the SAME administration rule above - do NOT default them to proctored. In an in-person course assume "in_person"; in an online/hybrid course assume "online_self" (unproctored) unless a webcam tool is named for that specific quiz/test (then "online_proctored") or the syllabus says it is held in person (then "in_person").

- * If a single bucket lists sub-components with individual weights, expand it into separate entries (do NOT use indeterminate). **EITHER/OR TRACKS: if the syllabus offers alternative assignment tracks of which a student completes only one (e.g. "Option A ... OR Option B ..."), code ONE track only.** Listing both inflates the denominator with work no student actually does. If the tracks carry the same total weight, either one is acceptable.

- * MULTI-POPULATION SYLLABI - if a syllabus has separate grading schemes for different student populations (e.g., undergraduate vs. graduate in the same course), use the REGISTERED COURSE NUMBER if provided (100-499 = undergraduate scheme, 500+ = graduate scheme); otherwise use the course number visible in the syllabus header/title. Do NOT merge schemes. Include graduate-only extra assignments only when analyzing the graduate population.

4c. "task_type" - WHAT KIND OF WORK the student produces. This is a SEPARATE question from 4a and 4b. 4a/4b say WHERE the work happens and whether anyone watches; 4c says WHAT the work IS. Never let one answer change the other: an in-class handwritten essay exam is still "written_prose", and a take-home multiple-choice quiz is still "objective_items".

Choose exactly one:

- * "written_prose": the student produces continuous prose. Essays, papers, memos, reports, reflections, journals, blog or discussion posts, literature reviews, annotated bibliographies, case write-ups, proposals, scripts, TRANSLATIONS, and free-text short-answer or essay exam questions.
- * "code_or_data": the student writes or runs code, scripts, queries or a data analysis. Programming assignments, notebooks, SQL, MATLAB, R, Stata, Python, spreadsheet model building, simulation work.
- * "quantitative_problem_set": the student solves set problems with numbers, formulas or proofs. Problem sets, exercise sets, derivations, proofs, publisher-platform drill (McGraw Hill Connect, Cengage, WileyPLUS, MyLab, Aplia, WebAssign, Quest, zyBooks), and quantitative exam questions.
- * "objective_items": the student picks or supplies short factual answers. Multiple choice, true/false, matching, fill-in-the-blank, vocabulary or grammar drill, listening-comprehension items, reading-check quizzes.
- * "oral_or_performance": the student speaks or performs. Presentations, talks, recitals, debates, recorded oral tasks, conversation or interview tasks, critiques, live skill demonstrations.
- * "studio_lab_physical": the student makes or does something physical. Studio and design work, bench or wet-lab work, clinical or field practice, fieldwork, physical performance and skills practice.
- * "participation": attendance, in-class discussion, engagement, peer review, and other credit for showing up or taking part rather than for a produced artefact.
- * "unspecified": the syllabus does not say what the work consists of.

>>> "unspecified" IS A CORRECT ANSWER, NOT A FAILURE. Use it whenever the document does not describe the content of the component. Do NOT infer the content from the department, the course level, or what a course like this usually contains. A component named only "Exam", "Midterm", "Homework", "Assignments" or "Quizzes", with nothing anywhere in the syllabus describing what is on it, is "unspecified".

>>> AN EXAM IS NOT A TASK TYPE. "Exam", "midterm", "final", "test" and "quiz" describe a format, not the work. Code what is ON it. "Final exam: three essay questions" is "written_prose". "Midterm: multiple choice" is "objective_items". "Final: derivations, show your work" is "quantitative_problem_set". If the syllabus never says what is on it, it is "unspecified".

>>> EVIDENCE MUST DESCRIBE **THIS** COMPONENT. This is the rule that is most often broken, so read it twice. Sitting in the same grading table is NOT evidence. If the syllabus says "Quizzes will contain multiple-choice and fill-in-the-blank questions" and says nothing at all about the Exams, then the QUIZZES are "objective_items" and the EXAMS are "unspecified". Do not carry a description across from a differently named component, in either direction, even when the two obviously belong to the same course and even when the

inference feels safe. If the only thing you can find describes some other component, the answer is "unspecified".

>>> NEVER READ TASK CONTENT OUT OF POLICY BOILERPLATE. Sections on academic integrity, scholastic dishonesty, sharing of course materials, plagiarism, copyright, Title IX, disability accommodation, counselling, religious holy days, campus carry, or emergency evacuation are boilerplate. They list many kinds of work in passing and describe none of them. In particular the sentence "no materials used in this class, including ... assessments (quizzes, exams, papers, projects, homework assignments), in-class materials, review sheets, and additional problem sets, may be shared online ..." is boilerplate and is NEVER evidence for any task_type.

>>> EVIDENCE THAT DOES COUNT, strongest first:

1. a description of the component itself, wherever it appears in the document;
2. the weekly schedule or course calendar, where it names that component;
3. required materials and equipment, when they are tied to that component ("a financial calculator is required for the exams", "bring a blue book");
4. the course learning objectives, when the course has one clearly dominant kind of graded work.

>>> A COURSE FLAG IS A LAST RESORT, NOT ORDINARY EVIDENCE. A stated UT course flag (Writing Flag, Quantitative Reasoning Flag) describes the COURSE, not any one component. Use it ONLY when BOTH of these hold:

- (a) the component has no description anywhere in the document, AND
- (b) the course has one clearly dominant kind of graded work.

When you use it, set evidence_location to "course_flag" so the choice can be audited and removed later. **If you are choosing between a course flag and "unspecified", choose "unspecified".** A flag must never be used to type more than a couple of components in one document: if most of the grading table has no description of its own, the honest answer for those components is "unspecified", not the flag applied over and over.

>>> LANGUAGE COURSES USE THE SAME EIGHT CATEGORIES. A composition or dialogue journal is "written_prose". A recorded oral task, conversation or interview is "oral_or_performance". A vocabulary or grammar quiz, or a listening exercise, is "objective_items". A translation exercise is "written_prose". Mixed online workbook practice that the syllabus describes as targeting reading, listening, writing and speaking together is "objective_items". Do not invent a language category.

4d. "task_evidence" – the VERBATIM words from the syllabus that justify 4c, at most 25 words. **COPY, DO NOT COMPOSE.** Do not merge two sentences, do not supply a missing subject, do not tidy the grammar, do not turn a phrase into a full sentence. If you cannot find a run of words in the document that says what the component consists of, the task_type is "unspecified" and this field is exactly "none". Never quote boilerplate.

NOTE ON PDF TEXT: extraction sometimes splits a hyphenated word across a space ("multiple- choice") or inserts page furniture inside a sentence ("Connect to 6 | P a g e allow you"). Copy what is actually there. Do not repair it, and do not let a broken word tempt you into rewriting the quote.

4e. "evidence_location" – where those words came from. One of:

"component_description", "course_objectives", "schedule", "required_materials", "course_flag", or "none" (use "none" when task_type is "unspecified").

5. "ai_policy": "prohibited", "conditional", "permissive", "unclear", or "no" Read the syllabus carefully for any mention of Artificial Intelligence, ChatGPT, LLMs, Copilot, or generative tools.

- * "prohibited": AI tools are explicitly banned or not allowed for course work.
- * "conditional": AI use is allowed only under specific conditions (e.g., permitted for brainstorming but not writing, allowed with citation, or restricted to certain assignments).
- * "permissive": AI tools are explicitly allowed or integrated into the coursework without significant restrictions.
- * "unclear": AI tools are mentioned but the policy is vague or ambiguous.
- * "no": No mention of AI tools anywhere in the syllabus.

6. "grading_basis": "letter", "pass_fail", "optional pass fail", or "unknown"

* "letter": course uses standard letter grades. This is the DEFAULT for ordinary university courses. Return "letter" whenever a letter-grade cutoff table is shown OR a weighted grade

breakdown (components in percent or points) is present and the course is not stated to be pass/fail-only. Do NOT require an explicit cutoff table – a percentage/points breakdown is sufficient evidence of letter grading.

- * "pass_fail": course is offered ONLY on a pass/fail basis. Synonyms: "credit/no credit", "CR/NC", "satisfactory/unsatisfactory", "S/U", "P/NP", "P/F". Look for explicit statements such as "This course is graded on a credit/no credit basis only" or "Grading basis: pass/fail".

- * "optional_pass_fail": course is letter-graded by default but the SYLLABUS ITSELF states students may elect pass/fail for THIS course. Signals: "students may elect P/F for this course", "this course may be taken for a letter grade or credit/no credit". DO NOT return "optional_pass_fail" based on the generic, university-wide pass/fail-election policy boilerplate that appears in most syllabi (e.g., a general paragraph about the registrar's Q-drop/pass-fail deadlines that applies to all courses) – that is just "letter".

- * "unknown": use ONLY when the syllabus shows NO grade breakdown, NO cutoff table, and NO grading-basis statement at all (genuinely silent). If any weighted breakdown exists, prefer "letter".

E.2 CV Coding Prompt

The system message is the short instruction below, and the user message is the template below, in which <CV text> is replaced by the text of one CV. When a surname is on file for the CV, the line "EXPECTED INSTRUCTOR SURNAME: <surname>" is placed before the text as a hint.

System message

You are a research assistant extracting factual career information from academic CVs. You must respond ONLY with a valid JSON object – no explanation, no markdown, no extra text.

User message template

First, decide what this document is. If it is NOT a CV or resume (for example a course syllabus, a reading list, a narrative biography with no dated entries, a cover letter, or unreadable OCR noise), still fill in what you can and set "document_type" accordingly. Report only what the document actually says.

NOTES ON THE INPUT:

- * A line "EXPECTED INSTRUCTOR SURNAME: <name>" may appear before the text. It is the surname on file for this document. It is a HINT ONLY. If the document clearly belongs to a different person, report the name you actually see in "name_on_document" – do not force a match.
- * Long CVs may have their later pages omitted, marked "[... MIDDLE AND LATER PAGES OF THE CV OMITTED ...]". Base your answers only on the visible text.
- * Years must be plain 4-digit integers. If no year is given for an item, use 0. Never invent or estimate a year.

Return a JSON object with exactly these seven fields (document_type, name_on_document, degrees, ut_austin_appointments, ut_austin_mentioned, earliest_ut_austin_activity_year, latest_year_anywhere):

1. "document_type": "cv", "resume", "bio_narrative", "syllabus", "other", or "unreadable"
 - * "cv": an academic curriculum vitae (education, appointments, publications).
 - * "resume": an industry-style resume (jobs and skills, little or no publication list).
 - * "bio_narrative": prose biography with no structured dated sections.
 - * "syllabus": a course syllabus filed in the CV slot.
 - * "other" / "unreadable": anything else, or text too garbled to read.

2. "name_on_document": the person's name exactly as printed at the top of the document. Empty string if none is visible.

3. "degrees": a list of EVERY degree the document mentions, in the order they appear.

>>> GATE – "LISTED" IS NOT "AWARDED". Read this before filling the list.
A degree is "completed" ONLY if the document shows it as actually awarded.
Set "status" to "expected" or "in progress" – NOT "completed" – whenever you see:
"expected", "anticipated", "in progress", "projected graduation", "Ph.D. Candidate",
"Doctoral Student", "ABD", or a year in the future relative to the rest of the document.
STILL LIST the degree, with its year. Do not drop it and do not silently promote it.

For each degree record:

- * "level": "doctorate" (PhD, EdD, DPhil, ScD), "professional" (JD, MD, DDS, DVM, MBA when it is the terminal degree), "masters" (MA, MS, MSc, MFA, MEd, MSW, MLA, MPH), "bachelors" (BA, BS, BBA, BSc, LLB, or a UK honours degree), "associate", "certificate", or "other".
- * "credential": the credential string exactly as printed ("Ph.D.", "M.F.A.", "B.B.A.", "2:1 (Hons)").
- * "field": the subject as printed. Empty string if not stated.
- * "institution": the awarding institution exactly as printed. Do NOT normalize or abbreviate.
- * "year": the 4-digit year the degree was awarded, or 0 if no year is given.
- * "year_basis": "stated" if a single award year is printed; "range_end" if only an enrollment range is given (e.g. "1979-1984" -> year 1984, basis "range_end"); "absent" if no year appears at all (then year is 0).
- * "status": "completed", "in progress", "expected", or "unknown".
- * "evidence": the verbatim line the degree came from, at most 120 characters.

RULES:

- * A POSTDOC IS NOT A DEGREE. Postdoctoral fellowships, research fellowships and traineeships are often printed inside the Education section. Do not list them as degrees.
- * Certificates, licences and non-degree training are "certificate" (or "other"), never "masters" or "doctorate".
- * If two degrees are awarded in the same year, list both.
- * If the document lists degrees with no years anywhere, still list them with year 0 and year_basis "absent".

4. "ut_austin_appointments": a list of EVERY position held AT THE UNIVERSITY OF TEXAS AT AUSTIN.

>>> GATE 1 - UT AUSTIN JOB TITLES. Read this before classifying anything.
 At UT Austin, "Assistant Instructor" (often abbreviated "AI") is a GRADUATE-STUDENT title, NOT a faculty position. So are "Teaching Assistant", "Associate Instructor", "Graduate Research Assistant", "Graduate Assistant", "Assistant Instructor of Record".
 -> role_class "graduate_student_instructor".
 By contrast, "Assistant Professor of Instruction" IS a faculty position.
 -> role_class "faculty_teaching".
 These two titles look alike and mean opposite things. Read the title carefully.

>>> GATE 2 - WHICH UNIVERSITY OF TEXAS. Include a position ONLY if it is at The University of Texas at AUSTIN. The following are DIFFERENT institutions and must be EXCLUDED from this list: UT Arlington, UT Dallas, UT San Antonio, UT El Paso, UT Pan American, UT Rio Grande Valley, UT Tyler, UT Permian Basin, UT Medical Branch, UT Southwestern, UT Health, Texas State University (formerly Southwest Texas State), Texas A&M, and Austin Community College. "University of Texas" with no campus named counts as Austin ONLY if the document elsewhere makes the Austin campus clear (an Austin address, a UT Austin department, utexas.edu).

For each appointment record:

- * "title": the title exactly as printed.
- * "unit": the department, school, centre or lab, as printed. Empty string if none.
- * "role_class": one of
 - "faculty_professorial": Assistant Professor, Associate Professor, Professor, Distinguished/Endowed Professor, Chair.
 - "faculty_teaching": Lecturer, Senior Lecturer, Instructor (non-student), Adjunct, Clinical Professor, Professor of Practice, Assistant Professor of Instruction, Visiting Assistant Professor teaching courses.
 - "graduate_student_instructor": see GATE 1.
 - "postdoc": Postdoctoral Fellow/Researcher/Associate, named postdoctoral fellowships.
 - "staff_administrative": Research Scientist, Engineering Scientist, Director, Associate Director, Coordinator, Program Manager, Advisor - non-teaching staff roles.
 - "visiting": Visiting Scholar, Visiting Researcher, Visiting Fellow, Guest Lecturer, Guest Critic, Visiting Design Juror, invited speaker - short-term or occasional roles that are not a standing appointment. A title that is BOTH visiting and a standing teaching post ("Visiting Assistant Professor", "Visiting Lecturer" for a full term) is "faculty_teaching", not "visiting".
 - "unknown": a UT Austin position whose nature is unclear.
- * "start_year": the 4-digit year the position began, or 0 if not stated.
- * "end_year": the 4-digit year it ended, or 0 if it is current or open-ended.
- * "is_current": true if the position is shown as ongoing ("present", "current", "-").
- * "evidence": the verbatim line, at most 120 characters.

>>> GATE 3 - WHERE THE YEARS SIT ON THE LINE. Read this before reading any dates.
 The date column is very often RIGHT-ALIGNED AT THE END of the line, not at the start:
 Associate Professor, Organizational Communication & Technology 2007 - present

Assistant Professor, Organizational Communication & Technology 2001 - 2007
Take each position's years from ITS OWN line. Never carry a year down to the title on the next line or up to the title on the previous line. If a line holds a title but no year, its start_year is 0 - do not borrow a neighbour's year to fill it. For a range on one line, start_year is the FIRST year and end_year is the SECOND ("2001 - 2007" -> start_year 2001, end_year 2007). Never report the end of a range as the start.
APPOINTMENT BLOCKS ARE OFTEN ORDERED BY RANK OR BY END DATE, NOT BY START DATE.
A promotion ladder is commonly printed most senior first: Professor, then Associate Professor, then Assistant Professor, so the EARLIEST position is often the LAST line of the block. List every line as printed; do not assume the first line is the earliest.

RULES:

- * LIST EVERY POSITION SEPARATELY, INCLUDING PROMOTIONS. "Assistant Professor 2002-2008" and "Associate Professor 2008-2015" are two entries. Do not merge them and do not try to work out which came first - just list them.

- * If a position has no start year, still list it with start_year 0.

- * If the person has never held a position at UT Austin, return an empty list [].

- * Do not put degrees in this list. A UT Austin degree is a degree, not an appointment.

5. "ut_austin_mentioned": true if The University of Texas at Austin appears ANYWHERE in the document for any reason (a degree, a job, a talk, a grant, an award, an address, an email). False if it never appears at all.

6. "earliest_ut_austin_activity_year": the EARLIEST 4-digit year attached to UT Austin WORK - a course taught at UT, a UT grant held, a UT award received as faculty or staff, a UT committee served on, a UT teaching-evaluation table row, a UT student advised. This is a fallback used when the appointment list has a gap, so cast a wide net over WORK.

>>> DO NOT USE ANY OF THESE: the year a degree was earned at UT Austin (being a STUDENT there is not activity - report it as a degree instead); the year of a conference held at UT Austin that the person merely attended or presented at; a UT Austin address or email with no year. If the only UT Austin years in the document come from a degree or from attending an event, return 0.

Use 0 if there is no qualifying year.

7. "latest_year_anywhere": the most recent 4-digit year mentioned anywhere in the document for any reason, including forthcoming work and future grant end dates. Use 0 if the document contains no years.

Return ONLY a JSON object. Examples:

Clean professor CV, promotions listed separately, PhD year from an enrollment range:

```
{ "document_type": "cv", "name_on_document": "[Name A]", "degrees": [{"level": "doctorate", "credential": "Ph.D.", "field": "History", "institution": "Cambridge University", "year": 1984, "year_basis": "range_end", "status": "completed", "evidence": "1979-1984 Ph.D., Cambridge University"}], "ut_austin_appointments": [{"title": "Assistant Professor", "unit": "Department of History", "role_class": "faculty_professorial", "start_year": 1984, "end_year": 1991, "is_current": false, "evidence": "1984-1991 Assistant Professor, Department of History, University of Texas at Austin"}, {"title": "Associate Professor", "unit": "Department of History", "role_class": "faculty_professorial", "start_year": 1991, "end_year": 0, "is_current": true, "evidence": "1991-present Associate Professor, Department of History"}], "ut_austin_mentioned": true, "earliest_ut_austin_activity_year": 1983, "latest_year_anywhere": 2023 }
```

Doctoral student who teaches: the in-progress PhD is NOT completed, and the "Assistant Instructor" title is a graduate-student role: { "document_type": "cv", "name_on_document": "[Name B]", "degrees": [{"level": "masters", "credential": "M.A.", "field": "Communication Studies", "institution": "The University of Alabama", "year": 2019, "year_basis": "stated", "status": "completed", "evidence": "December 2019 Master of Arts in Communication Studies"}, {"level": "doctorate", "credential": "Ph.D.", "field": "Communication Studies", "institution": "The University of Texas at Austin", "year": 2024, "year_basis": "stated", "status": "in_progress", "evidence": "Ph.D. Communication Studies, Projected Graduation May 2024"}], "ut_austin_appointments": [{"title": "Assistant Instructor", "unit": "Department of Communication Studies", "role_class": "graduate_student_instructor", "start_year": 2021, "end_year": 0, "is_current": true, "evidence": "Fall 2021 Assistant Instructor, Department of Communication Studies"}], "ut_austin_mentioned": true, "earliest_ut_austin_activity_year": 2020, "latest_year_anywhere": 2024 }

Appointment list has a gap but UT teaching evidence starts earlier - report both, do not reconcile them: { "document_type": "cv", "name_on_document": "[Name C]", "degrees": [{"level": "doctorate", "credential": "Ph.D.", "field": "Counseling and Psychological Services", "institution": "Georgia State University", "year": 1998, "year_basis": "stated", "status":

```
"completed", "evidence": "Ph.D. Georgia State University 1993-1998"}], "ut_austin_appointments": [{"title": "Associate Professor", "unit": "African and African Diaspora Studies", "role_class": "faculty_professorial", "start_year": 2010, "end_year": 2013, "is_current": false, "evidence": "University of Texas at Austin 2010-2013 Associate Professor"}], "ut_austin_mentioned": true, "earliest_ut_austin_activity_year": 2007, "latest_year_anywhere": 2027}
```

Degree from a DIFFERENT UT campus and no UT Austin job - the UT Arlington doctorate is still a degree, but there are no UT Austin appointments: {"document_type": "cv", "name_on_document": "[Name D]", "degrees": [{"level": "doctorate", "credential": "Ph.D.", "field": "Social Work", "institution": "University of Texas at Arlington", "year": 2009, "year_basis": "stated", "status": "completed", "evidence": "UNIVERSITY OF TEXAS AT ARLINGTON Doctor of Philosophy Graduation: May 2009"}], [{"level": "masters", "credential": "M.S.S.W.", "field": "Social Work", "institution": "University of Texas at Austin", "year": 2006, "year_basis": "stated", "status": "completed", "evidence": "UNIVERSITY OF TEXAS AT AUSTIN Master of Science in Social Work, 2006"}], "ut_austin_appointments": [], "ut_austin_mentioned": true, "earliest_ut_austin_activity_year": 0, "latest_year_anywhere": 2021}

Degrees listed with no years anywhere in the document: {"document_type": "cv", "name_on_document": "[Name E]", "degrees": [{"level": "doctorate", "credential": "Ph.D.", "field": "Mathematics", "institution": "University of California at Irvine", "year": 0, "year_basis": "absent", "status": "completed", "evidence": "Ph.D. (Mathematics) University of California at Irvine, California"}], [{"level": "masters", "credential": "M.A.", "field": "Mathematics", "institution": "University of California at Irvine", "year": 0, "year_basis": "absent", "status": "completed", "evidence": "M.A. (Mathematics) University of California at Irvine"}], "ut_austin_appointments": [{"title": "Assistant Professor", "unit": "Departments of Mathematics and General Business", "role_class": "faculty_professorial", "start_year": 1977, "end_year": 1980, "is_current": false, "evidence": "1977-1980 Assistant Professor, Departments of Mathematics and General Business"}], "ut_austin_mentioned": true, "earliest_ut_austin_activity_year": 1977, "latest_year_anywhere": 2025}

Right-aligned date column, promotion ladder printed MOST SENIOR FIRST - the earliest appointment is the LAST line, and each title takes the dates from its own line: {"document_type": "cv", "name_on_document": "[Name F]", "degrees": [{"level": "doctorate", "credential": "Ph.D.", "field": "Communication", "institution": "University of California, Santa Barbara", "year": 2002, "year_basis": "stated", "status": "completed", "evidence": "2002 Ph.D., Communication, University of California, Santa Barbara"}], "ut_austin_appointments": [{"title": "Area Head, Organizational Communication & Technology", "unit": "Moody College of Communication", "role_class": "staff_administrative", "start_year": 2016, "end_year": 2019, "is_current": false, "evidence": "Area Head, Organizational Communication & Technology 2016 - 2019"}, {"title": "Associate Professor", "unit": "Organizational Communication & Technology", "role_class": "faculty_professorial", "start_year": 2007, "end_year": 0, "is_current": true, "evidence": "Associate Professor, Organizational Communication & Technology 2007 - present"}, {"title": "Assistant Professor", "unit": "Organizational Communication & Technology", "role_class": "faculty_professorial", "start_year": 2001, "end_year": 2007, "is_current": false, "evidence": "Assistant Professor, Organizational Communication & Technology 2001 - 2007"}], "ut_austin_mentioned": true, "earliest_ut_austin_activity_year": 2004, "latest_year_anywhere": 2025}

The file is not a CV at all (a course syllabus was filed in the CV slot): {"document_type": "syllabus", "name_on_document": "[Name G]", "degrees": [], "ut_austin_appointments": [], "ut_austin_mentioned": true, "earliest_ut_austin_activity_year": 2019, "latest_year_anywhere": 2019}

```
CV TEXT
<CV text>
END
```

References

- Akter, Syeda Nahida, et al. 2023. "An In-depth Look at Gemini's Language Abilities." arXiv 2312.11444.
- Center for AI Safety and Scale AI. 2026. "Humanity's Last Exam." AI Dashboard, lastexam.ai. Accessed September 30, 2026.
- 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.
- Dua, Dheeru, et al. 2019. "DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*: 2368-2378.
- Gemini Team, Google. 2023. "Gemini: A Family of Highly Capable Multimodal Models." arXiv 2312.11805.
- 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. 2023. "GPT-4 Technical Report." arXiv 2303.08774.
- OpenAI. 2024a. "simple-evals." GitHub repository, github.com/openai/simple-evals.
- Rein, David, et al. 2023. "GPQA: A Graduate-Level Google-Proof Q&A Benchmark." arXiv 2311.12022.
- Shi, Freda, et al. 2022. "Language Models Are Multilingual Chain-of-Thought Reasoners." arXiv 2210.03057.
- Wei, Jason, et al. 2024. "Measuring Short-Form Factuality in Large Language Models." arXiv 2411.04368.