Can AI Ace Your Exam? A Penn Law Experiment

R. Polk Wagner[†]

University of Pennsylvania Carey Law School

Early Draft — June 2026

Not for publication or circulation

Updated drafts, abstracts, slides, and other project materials live at
`ai-teaching-lab.org/projects/exam-taker`

Abstract

I report a controlled experiment that ran a frontier large language model through eleven real University of Pennsylvania Carey Law School final exams in Spring 2026 and sent the machine's answers out for blind grading alongside the work of enrolled students. Each exam was answered twice: once with no course materials in the model's context (the "no-materials" arm, relying on the model's training alone) and once with the course syllabus and an outline supplied (the "with-materials" arm). Three findings emerge. First, the model is broadly competitive: across the ten exams returned with grades, it passed every one, sat at or above the class median on every one, and reached the top decile on seven. On two exams, both scored on an open, non-saturated scale, it outscored every enrolled student. Second, the benefit of supplying course materials is real but concentrated and uneven. It is large on multiple-choice components, small to absent on essays, and on one exam the materials *helped* the machine's multiple-choice score while *hurting* its essay—so the gain vanishes when the two are averaged into a single number. The subject of the exam matters as much as its format. Third, faculty graders detected the AI submissions at a high rate, but mostly through signals that are easy to remove: formatting artifacts, identification-number placement, even a reversed quotation-mark glyph. The detection that survives those fixes is content-based—two independently generated answers that read as though written by the same hand, and a voice that does not match time-pressured student work. One grader was confident enough in detection to decline to grade the AI answers at all. I read these results less as a verdict than as a measurement problem: when a model that never took the course can top the curve, it is no longer clear that the time-limited, high-stakes final is measuring what we have long assumed

[†]Michael A. Fitts Professor of Law at the University of Pennsylvania Carey Law School. Special thanks to Tulio Tagliaferri (Penn Carey Law J.D. Class of 2027) for his substantial contributions to this project; to Claire Wallace, Senior Associate Dean and Penn Carey Law Registrar, and her Academic Affairs team; and to the ten faculty who volunteered their exams for this study. This is an early working draft (June 2026) reporting Wave 1 (Spring 2026) results. Faculty and students are anonymized in this draft.

it measures—mastery of *this* course—rather than fluency in a genre the machine commands. That genre was always a contrivance built for grading, not a faithful model of practice, so the machine's success says less about lawyering than the headline suggests; but it says a good deal about how much weight the instrument can still bear. The urgent ask the experiment surfaces is intentionality: that faculty be clear about what each assessment is meant to measure, and recognize that AI is widening an already-real gap between the law-school exam and the practice it stands in for. The responses will range—from rethinking assessment top to bottom, to narrower adjustments, to simply naming the mismatch and grading with it in view—and reasonable colleagues will differ.

I. Introduction

Can an artificial intelligence pass your exam? Not a practice question written for a benchmark, and not a released bar exam scored against a public answer key, but the actual final you wrote for the students sitting in your course this term, graded the way you grade everyone else, mixed into the same stack.

That is the question this project set out to answer, and the answer is now mostly in. Over the Spring 2026 examination period at the University of Pennsylvania Carey Law School, a frontier large language model sat for eleven final exams—Federal Income Tax, Remedies, two sections of Constitutional Law, two of Criminal Law, Intellectual Property, Legislation, National Security Law, International Law, and a seminar on criminal justice and *The Wire*. The machine's answers were packaged to look like ordinary student submissions, given reserved examination numbers the registrar guaranteed would not collide with any real student, and dropped into the grading pile. The participating faculty graded blind, on their own rubrics, and reported back.

The headline is easy to state and hard to dismiss: the model did well. On the ten exams returned with grades, it earned a passing grade on every one, landed at or above the class median on every one, and reached the top ten percent of the class on seven. On the Legislation exam it scored 99 and 100 out of 100—above all eighty-four enrolled students. One constitutional law grader called the machine's answer "the best of the lot" in a class of ninety-two. A criminal-law instructor, shown the multiple-choice results, watched the machine move from the class average to the top of the class when course materials were added to its context.

And yet the interesting findings are not in the headline. They are in three places the headline obscures, and this Article is organized around them.

First, what the machine can do, it can largely do without help. The experiment was built to measure one thing above all: whether attaching course materials to the model's context improves its performance. The intuition—shared by most faculty I described the project to—was that a model armed with the syllabus and a good outline would clearly outperform a model working from training alone. That intuition is only sometimes right, and the pattern is more revealing than a simple yes. On multiple-choice components the materials help a great deal, sometimes moving the model forty percentage points. On essays the materials barely move the needle, and on one exam they made the essay *worse*. Because most exams blend the two, the effect of materials on a student's bottom-line grade is frequently near zero—not because materials do nothing, but because they push the two halves of the exam in opposite directions. The model's essay competence, it turns out, comes mostly from training, not from the outline. The subject matters too: on constitutional law, a field the model handles fluently from training alone, the no-materials arm reached the top of the class without any course-specific grounding at all.

Second, the machine is competitive at the top of the class, not merely passing. "Can AI pass

law school" sets a low bar that the technology cleared some time ago. The sharper question is where in the distribution the machine lands when it is graded against real students on a real curve, and the answer is: near the top, consistently, and on the strength of training rather than retrieval. I report the standing three ways—pass, at-or-above-median, and top-decile—because each tells a different part of the story and because the framing one chooses changes the rhetorical stakes.

Third, faculty can tell—but mostly for reasons that will not last. I measured detection as carefully as the scores, and the result has two layers. Graders identified the AI submissions at a high rate. But when one separates *how* they did it, most of the early detections rode on artifacts that have nothing to do with the quality of the writing: the machine's pages were single-spaced where the proctoring software double-spaces student work; the font differed; the reserved examination numbers sat conspicuously at the end of an ID-sorted list; one answer rendered opening quotation marks backwards. Every one of these is a build-script bug, fixable before the next wave. What remains after the fixes is harder to engineer away and more interesting to think about: the two independently generated answers to the same exam often read as though one author wrote both, and the prose carries a thoroughness and polish that time-pressured student work, with its spelling slips and ragged edges, does not. Three different graders, on three different exams, independently described the two AI answers as sounding like "the same individual or entity." That is the durable signal, and one grader found it so unmistakable that he declined to grade at all, concerned that scoring work he could identify as machine-written would no longer be an unbiased exercise.

A quieter finding runs underneath these three. The machine tops curves built to reward coverage and polish, but on one exam the grader singled out a human answer that did things no machine answer in that course did: a mock judicial opinion built around a creative conceit, witty and structurally daring. It is a single observation, from one course and one reader, but it points past what the rubric scores. The machine reaches the ceiling of the competent, comprehensive answer; the heights above it, where originality and judgment live, it did not reach here. I return to this in Parts IV and VI, because it bears on what we are testing for.

A recent study points the same direction from a different angle. As this project was running, a Stanford-led team asked the tutoring-channel version of the question and found that law professors, grading blind, prefer AI answers to their colleagues'. Their result and this one reach a similar conclusion from opposite ends, preference in office hours and performance on the graded exam, and I situate the two in Part II.

These findings bear on a question the legal academy is having right now: what an exam still measures once a machine can pass it blind, and how much of what we assumed it measured it was measuring all along—a question sharpened by the fact that the exam was always a contrived proxy for lawyering, not the thing itself. I take that up in Part VI. But the contribution I want to foreground is methodological. The "AI passes the bar" genre has been criticized, fairly, for testing

models against released questions and public rubrics under conditions that flatter the machine.[1] This experiment was designed to remove that flattery: real exams the model had never seen, blind human grading on the instructor's own rubric, a real student curve as the comparator, and a pre-registered analysis plan locked before the bulk of the grades came back. The design is the point, and Part III sets it out in full.

A word on what this study is and is not. It is a pilot—Wave 1 of a planned multi-wave project. It uses a single model from a single vendor, a single run per exam cell, and a cohort of eleven exams assembled from faculty who volunteered. Those are real limitations, and I treat them as such in Part VII rather than burying them. They are also the appropriate posture for a pilot whose second purpose, alongside producing data, is to expose what a larger study should tighten. Part VIII describes how Wave 2 (Fall 2026) tightens it.

The Article proceeds as follows. Part II situates the study in the literature on AI and legal assessment and explains how its design differs. Part III describes the methodology in detail—how the AI answers were written, how isolation between conditions was enforced, how the answers were packaged for blind grading, and what was pre-registered. Part IV reports how the machine performed and dissects the materials effect. Part V reports whether and how faculty detected the AI submissions. Part VI discusses the implications. Part VII catalogs the limitations. Part VIII previews Wave 2. A short conclusion follows.

II. Background and Prior Work

A. The "AI passes the exam" literature

The claim that generative AI can pass professional examinations entered public consciousness through the bar exam. Early work reported that a GPT-class model scored in the lowest decile of a simulated multistate bar exam; within a year, a successor model was reported to pass a full simulated bar in the ninetieth percentile.[2] In parallel, a study of a model's performance across actual law-school course exams—taken under conditions the authors arranged with their own institution—reported passing but unremarkable grades, clustered near the bottom of the class.[3] These studies established both that the technology had crossed the passing threshold and that the

[1] For the leading example of the genre, see Daniel Martin Katz, Michael James Bommarito, Shang Gao & Pablo Arredondo, *GPT-4 Passes the Bar Exam*, 382 Phil. Transactions Royal Soc'y A 20230254 (2024), and for the principal methodological critique, Eric Martínez, *Re-evaluating GPT-4's Bar Exam Performance*, 33 Artificial Intelligence & L. 581 (2025). See *infra* Part II.A.

[2] Michael James Bommarito II & Daniel Martin Katz, *GPT Takes the Bar Exam* (Dec. 29, 2022) (unpublished manuscript), https://dx.doi.org/10.2139/ssrn.4314839 (GPT-3.5; passed the multiple-choice components on evidence and torts but failed overall); Katz, *supra* note 1, at 20230254 (GPT-4; roughly top-decile result on a simulated Uniform Bar Exam).

[3] Jonathan H. Choi, Kristin E. Hickman, Amy B. Monahan & Daniel B. Schwarcz, *ChatGPT Goes to Law School*, 71 J. Legal Educ. 387 (2023) (vol. 71, no. 3, at 387–400). SSRN DOI 10.2139/ssrn.4335905.

legal academy would treat the milestone as significant.

And yet they also drew careful criticism. The most pointed argued that the headline ninetieth-percentile bar result rested on comparisons against unrepresentative populations and on scoring choices that inflated the estimate; measured against first-time test-takers, the same performance looked considerably more modest.[4] The broader methodological objection is one this project takes seriously: when a model is tested against released questions, public rubrics, or benchmark datasets that may overlap its training data, the result tells us less about capability on novel work than the framing suggests.

This study sits in that lineage but changes the test conditions in four ways. The exams are real, current, and unreleased—written for enrolled students in Spring 2026 and never published. The grading is done blind by the faculty member who wrote the exam, on that instructor's own rubric, with the AI answers mixed into the live grading stack. The comparator is the actual class curve, computed after excluding the AI submissions so the machine is measured against real students only. And the analysis plan was pre-registered before most of the grades returned. The goal is to measure capability under conditions that do not flatter the machine.

B. The materials manipulation

A second body of work concerns *retrieval-augmented* performance: whether giving a model relevant documents at inference time improves its output.[5] The legal-education question is the applied version—does a student-style "open book" of course materials help an AI the way it helps a student? This project makes that the central experimental manipulation. Every exam was answered twice under otherwise identical conditions: once with nothing but the model's training (the no-materials arm), once with the course syllabus and an outline supplied in context (the with-materials arm). The contrast isolates the value of course-specific grounding, holding the model, the prompt, and the exam constant. As Part IV shows, the answer is not uniform, and the non-uniformity is the finding.

C. Detection and the assessment-integrity debate

A third literature, growing quickly, concerns whether AI-generated text can be reliably detected—and the emerging consensus is skeptical, with automated detectors prone to both false positives and easy evasion.[6] Most of that work studies *automated* detectors. This project supplies something

[4] Martínez, *supra* note 1, at 581–604.

[5] The foundational treatment is Patrick Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, 33 Advances in Neural Info. Processing Sys. (NeurIPS) 9459 (2020). The retrieval-augmented-generation literature is large; the applied claim here needs only the basic premise that supplying relevant documents at inference time can change a model's output.

[6] See, e.g., Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut & Lorna Waddington, *Testing of Detection Tools for AI-Generated Text*, 19 Int'l J. Educ. Integrity

rarer: data on whether expert human readers—law professors grading their own exams—can identify machine-written answers, and on what cues they actually use. That evidence speaks directly to the assessment-integrity debate now unfolding in law schools, which has so far proceeded largely on intuition about what AI writing "sounds like." Part V reports what the cues were and separates the durable signals from the fixable artifacts.

D. A recent related study: preference in the tutoring channel

As this project was running, a Stanford-led team ran the tutoring-channel version of the same question and reported a result worth pausing on: graded blind, law professors prefer AI answers to their colleagues'.[7] Sixteen Contracts professors across fourteen schools made nearly three thousand forced-choice comparisons between anonymized short answers to office-hours questions, one written by a professor and one by a model, each time choosing which they would rather give a student. The models won roughly three-quarters of the matchups, were flagged as pedagogically "harmful" far less often than the human answers, and—on an expert-validated extension that used a language model as the judge—were led by Claude Opus 4.7, the model this study uses.

The two studies are independent but complementary, and the contrast sharpens what is distinctive here. That study measures *preference* in a no-stakes tutoring setting; this one measures *graded performance* on real exams, blind, against the class curve. That study deliberately withheld course-specific grounding, and found that its one retrieval-augmented system underperformed the stock model—the tutoring-side echo of a result in Part IV, that course materials add less to the machine's essays than intuition predicts. And it did not study detection, which is this Article's second contribution. Read together, the two suggest that frontier models now meet the professional standard in law from two directions, the office-hours answer and the graded exam—while leaving open the questions this study takes up: where in the class distribution the machine lands when the work is actually graded, whether course materials help, and whether the faculty can tell.[8]

---

art. 26 (2023) (commercial detectors neither accurate nor reliable, and biased toward classifying machine text as human); Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu & James Zou, *GPT Detectors Are Biased Against Non-Native English Writers*, 4 Patterns 100779 (2023) (detectors misclassify non-native human writing as AI).

[7] Alejandro Salinas et al., *Law Professors Prefer AI Over Peer Answers* (May 27, 2026), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6849678 (manuscript) (senior author Julian Nyarko, Stanford).

[8] That study positions its result against a more cautious recent literature on model reliability in law. See Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning & Daniel E. Ho, *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*, 22 J. Empirical Legal Stud. 216 (2025); Matthew Dahl, Varun Magesh, Mirac Suzgun & Daniel E. Ho, *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models*, 16 J. Legal Analysis 64 (2024). On the tutoring side, see Greg Kestin, Kelly Miller, Anna Klales, Timothy Milbourne & Gregorio Ponti, *AI Tutoring Outperforms In-Class Active Learning*, 15 Sci. Rep. 17458 (2025) (randomized trial; larger learning gains from an AI tutor than from in-class active learning).

III. Methodology

This Part describes how the AI answers were produced, how the two conditions were kept from contaminating each other, how the answers were packaged so they could be graded blind, and what was fixed in advance by pre-registration. The full operating record lives in the project's methodology files; what follows is the account a reader needs to evaluate the results.[9]

A. Design

The experiment is a two-by-two between-runs design, replicated across exams (Figure 1). The two factors are the *question* (each exam's essay questions and multiple-choice block are treated as separate units) and the *materials condition* (no-materials versus with-materials). For a two-essay exam this yields four cells: question one with and without materials, question two with and without materials. Exams with a multiple-choice component add cells for that block. One exam (Remedies) added a third materials condition at the instructor's request—the six-page reference sheet students were permitted to bring into the room—analyzed only within that exam and not pooled with the rest.

"Between-runs" is the load-bearing choice. Each cell was generated by a fresh, isolated model invocation that could not see any other cell's prompt, reasoning, or output. The alternative—generating both arms of a question in one session—would let the with-materials arm's reading of the outline bleed into the no-materials arm through shared context, collapsing the very contrast the experiment exists to measure. Isolation removes that pathway. Its cost is that a single run per cell cannot estimate run-to-run variance; I return to this in Part VII.

B. The model

Every cell across every exam used the same model: Claude Opus 4.7 (one-million-token context), at default sampling, accessed through a command-line harness. A single model from a single vendor is a deliberate Wave 1 scope decision, not an oversight—the point of the pilot is to stabilize the protocol within one model before adding the complications of cross-vendor comparison. Multi-vendor coverage is a Wave 2 priority (Part VIII). The single-vendor choice is also a limitation on generalizability, flagged as such in Part VII.

C. How the AI answers were written

Each answer was produced by a structured prompt built from a fixed template, with only the exam-specific fact pattern swapped in. The template is organized into labeled blocks: a *role* block casting

[9] The project's methodology files—a project-level record, per-exam records for each of the eleven exams, a decision log, a lessons-learned log, and the pre-registered analysis plan—are maintained by the AI Teaching Lab and are the source for the figures in this Article; they are on file with the author. An anonymized replication dataset (per-exam scores, z-scores, and ranks, without identifying information) will be posted publicly upon publication.

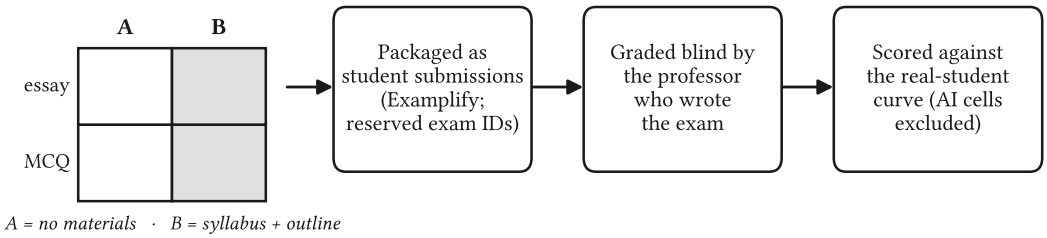

Figure 1: The experiment. Each of eleven exams was answered twice—once with no course materials, once with the syllabus and an outline—in isolated model runs, packaged to look like ordinary student submissions, graded blind by the professor who wrote the exam, and scored against the real-student curve with the AI answers removed.

the model as a strong Penn Carey Law student sitting the exam; a *task* block stating the objective; a *success-criteria* block describing a good answer; a *format* block imposing the output constraints (continuous prose, no headings, a word ceiling matched to the real exam's instructions); a *style* block; a *citation-honesty* block; an *execute* block instructing the model to return only the answer, with no preamble or commentary; and the *exam-question* block carrying the fact pattern.

Two blocks deserve emphasis because they shape what the grader saw. The *style* block bans a list of words and constructions that signal machine authorship—the throat-clearing transitions and intensifiers ("notably," "importantly," "furthermore," "this raises important questions") that models reach for under uncertainty—and prohibits the em-dash, a character LLMs overuse. The intent was to suppress the most obvious prose tells so that any detection would have to rest on something deeper than surface style. As Part V shows, this largely worked: where graders caught the machine on content, it was not because the prose was littered with "moreover." The *citation-honesty* block forbids invented authorities outright. The model was instructed that if it was unsure of a case or statute, it should state the rule without attributing it to a named source rather than fabricate a citation. The risk this guards against is well documented: general-purpose models hallucinate legal authorities at high rates, and even commercial legal-research tools that advertise otherwise still do.[10] Across the first-pass review of every cell, no fabricated authority survived into a delivered answer.

[10] See Dahl, *supra* note 8 (general-purpose LLMs); Magesh, *supra* note 8 (commercial legal-research tools).

D. Materials, and the line the materials could not cross

In the with-materials arm, the model received two kinds of document: the course syllabus and one or more outlines. Outlines came from one of two sources. Where a prior-year student outline was available, that was used—approximating "what a well-prepared student studied from." Where the operator had richer underlying material (lecture transcripts, the casebook, slides), an outline was synthesized for the model by a separate pipeline of isolated agents and then audited—approximating "what optimized AI-tailored preparation provides." Which source each exam used is recorded in the project's materials inventory. This heterogeneity—the with-materials arm is not a uniform dose across exams—is a known limitation (Part VII) and a principal driver of the Wave 2 design.

One constraint governed the materials absolutely: nothing supplied to the model could reveal the actual exam. Not the questions, not the day-of instructions, not the exact number of multiple-choice items or the precise word limit, not the structural breakdown unique to this term's exam. General format language that genuinely appeared in a pre-semester syllabus ("twenty to thirty multiple-choice questions plus one or two essays") was in bounds; anything that could only be known by reading the actual exam was not. The synthesis agents that built AI outlines were given file lists scoped to course content only and were forbidden from reading the exam file. Before any with-materials cell was dispatched, the operator confirmed that no file in the materials folder referenced this term's exam structure. The reason is contamination control: if the materials leak the exam's shape, the condition stops measuring "what a student studied with" and starts measuring "what a student studied with, plus a hint"—which would inflate the apparent value of materials and confound the contrast.

E. Capture, audit, revise

The protocol ran in three stages, and the order matters for the integrity of the record. First, *verbatim capture*: each cell's output was saved to disk exactly as the model returned it, including any protocol violations—preambles, em-dashes, stray formatting. The unedited output is part of the research record. Second, a *formal-compliance audit*: a deterministic set of shell checks counted words, searched for the banned em-dash and the banned phrases, and flagged headers, bullets, and italicized case names. The audit checks form, not substance; it does not assess whether the law is right. Third, *minimal revision* where the audit found violations: the operator (not a fresh model) made surgical edits—removing a preamble, replacing an em-dash with a comma, trimming an over-length answer by cutting from its weakest point. New analysis was never added; identified issues were never removed; citations were never altered. Both the first-pass and revised files are retained, and the difference is documented per cell. Cleaning up the output silently on the first pass would have destroyed the evidence of how the model actually behaves, which is itself a finding—the with-materials arm, for instance, had a persistent habit of opening with a narrative preamble that the no-materials arm did not, and that pattern is now a Wave 2 prompt-design target.

F. Packaging and blind delivery

To be graded blind, the AI answers had to look like student submissions. Penn Carey Law administers in-class exams through Examplify and exports them in a specific PDF format; a build script renders the AI answers into that format. Multiple-choice answers were packaged into the spreadsheet format the registrar forwards to instructors, with the AI submissions appearing on their own rows in the same workbook as the real students—so that an AI identification number would not stand out by having an essay but no multiple-choice entry, or vice versa. The registrar reserved three examination numbers for the whole term—one for each materials condition—and guaranteed they would not collide with any real student in any Spring 2026 exam.

This packaging is where Wave 1's most consequential blind-grading leakage occurred, and I report it plainly because it shapes the detection results in Part V. The build script's defaults did not perfectly match the Examplify student export: the AI pages were single-spaced where student exports are double-spaced, the body font differed, and the per-question page annotations Examplify adds were missing. The reserved examination numbers, all just above the real student range and clustered within a few dozen of each other, sat conspicuously at the end of any list sorted by ID. None of these is a property of the writing. All of them gave graders a way to identify the AI submissions that has nothing to do with whether a machine can write a good exam answer. Part VIII describes the fixes.

G. Pre-registration

The analysis plan was pre-registered on May 7, 2026—after all cells were generated and packaged, but before the bulk of the grades returned (three of the eventual ten numeric returns were in hand). The pre-registration fixed the primary contrast (no-materials versus with-materials), the outcome variables (raw score, z-score against the real-student-only distribution, percentile and rank, and faculty AI-flag), the pooling method (random-effects across exams, on the standardized effect), the format stratification and its hypothesis (that the materials effect would be larger on multiple-choice-bearing exams), and three operational definitions of "the AI passed"—competitive, at-or-above-median, and wins. It also fixed the multiple-comparisons posture: descriptive effect sizes with confidence intervals are primary; formal significance tests are secondary and corrected across the pre-registered family. Locking these choices before unblinding is what separates a measured result from a story told after the fact, and any departure from the plan is logged with its rationale.

IV. Results: How the Machine Performed

I report the performance results in three movements. First, the bottom line—where the machine landed against real students, under the three pre-registered framings. Second, the materials effect, which is the experiment's central manipulation and more interesting than a single number. Third, the finding that ties the two together: the machine's strength is heavily a function of *what kind* of

question it faces, not just whether it has the course materials in hand.

Three reading notes. Throughout, I standardize each score against the real-student-only distribution—the class curve recomputed with the AI submissions removed—and report it as a z-score (standard deviations above or below the real-student mean) so that results are comparable across exams scored on different scales. And on one exam (a Criminal Law section) the essays were not returned for grading; its multiple-choice result is reported, its composite is not, and nothing in the analysis turns on it. And the pass, median, and top-decile framings below credit the better of an exam's two arms—the materials condition that scored higher—with the per-arm figures shown in the Appendix.

A. The bottom line: competitive at the top

Of the eleven exams, ten have returned faculty judgments of the essay or composite work. (The eleventh, a Criminal Law section, returned only its multiple-choice result; its essays were never graded.) On all ten, the machine earned a passing grade. On all ten, at least one arm sat at or above the class median. On seven of the ten, at least one arm reached the top decile of the class (Figure 2).

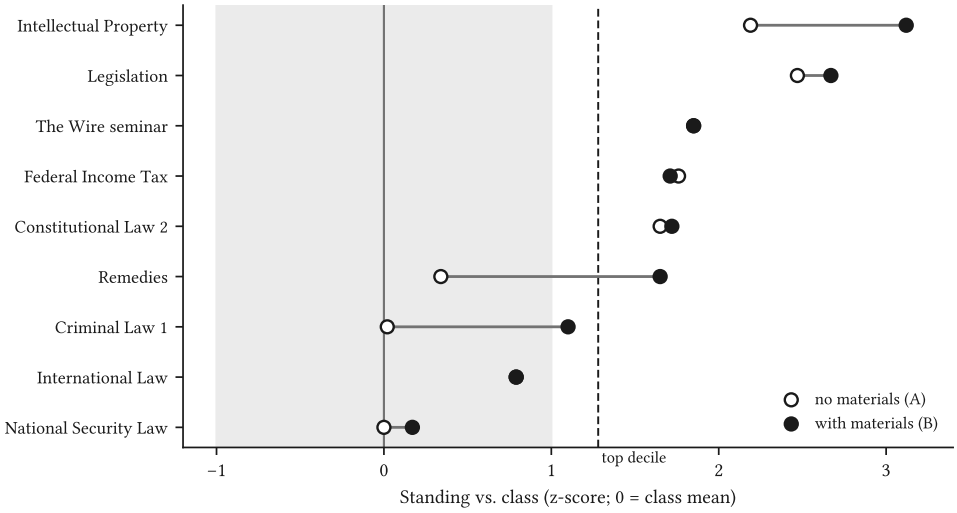

Figure 2: Where the machine landed. Each exam's AI answer as a z-score against the real-student-only distribution: 0 is the class mean, the shaded band is ±1 standard deviation, and the dashed line marks the top decile. Open circles are the no-materials arm, filled circles the with-materials arm. The nine exams with numeric composite scores are shown; Constitutional Law 1 returned only a qualitative judgment and Criminal Law 2 only its multiple-choice result.

The standing is not marginal. On the Legislation exam, the no-materials arm scored 99 out of 100 and the with-materials arm a perfect 100—both above every one of the eighty-four enrolled

students, whose top score was 98.[11] On the Intellectual Property exam the machine took the top composite score in a class of seventy-seven, in both arms. On the *Wire* seminar both arms reached the rubric ceiling, 40 out of 40, which the top real student did not. On one Constitutional Law exam the machine placed second and fifth of ninety-one on the combined score; on the Remedies exam the with-materials arm ranked fourth of forty-one. The weakest *combined* showing among the graded exams was still respectable, roughly the sixty-fifth percentile on National Security Law; the weakest among the essay-only exams was rank nine of fifty-two, about the eighty-fourth percentile, on the International Law take-home.

Two graders captured what the numbers compress. The constitutional-law instructor who declined to grade the AI answers—more on that in Part V—described the with-materials answer as "superb: comprehensive, rigorous, sophisticated, clearly and punchily written," and "clearly the best of the lot" in a field of ninety-two real exams. The international-law grader, whose exam produced the machine's weakest standing among the essay-only exams, still placed both arms in the low-A range and offered the most useful one-sentence account of the machine's profile in the whole cohort: "very good at matching facts to issues, very good at not missing issues, a little weaker at making more subtle arguments and drawing connections. Did better on the more straightforward questions." That profile—strong coverage, weaker synthesis—recurs across the cohort and explains why the machine tops some curves and merely beats the median on others.

Against the pre-registered framings, then: "AI is competitive" (a passing grade on at least eighty percent of exams) holds on ten of ten. "AI is at or above median" holds on ten of ten. "AI wins" (top-decile on at least half the exams) holds on seven of ten. By the project's own pre-committed standards, the machine did not merely pass. It competed at the top of the class.

B. The materials effect lives in the components, not the composite

Now the central question: does supplying course materials help? The pre-registered hypothesis, and the prior of nearly every faculty member I spoke with, was that it should—and that the help should be larger on multiple-choice questions, where a fact recalled is a point earned, than on essays, where the model's baseline competence is already high. The data half-confirm this and half-complicate it, in a way that turns out to be the most important analytic point in the Article.

Start with multiple-choice, where the effect is unambiguous and large. Across the six exams with a multiple-choice component, adding materials raised the machine's score by an average of roughly fourteen to fifteen percentage points, and on the two hardest multiple-choice instruments the jump was enormous: on one Criminal Law exam, from 45% (the class average) to 85% (tying the top of the class)—a swing of forty points, or more than two and a half standard deviations; on

[11] Unless otherwise noted, every figure in this Part traces to the project's scoring record; the per-exam detail is in the Appendix and the underlying `scoring-results.csv`. Ranks are stated against the real-student-only distribution except where a "full-set" rank (including the two AI submissions) is noted.

National Security Law, from 50% to 73.5%, about twenty-three points. The materials do for the machine on multiple-choice exactly what an outline does for a student: they supply the specific doctrinal facts that recall-style questions reward. (See Figure 3.)

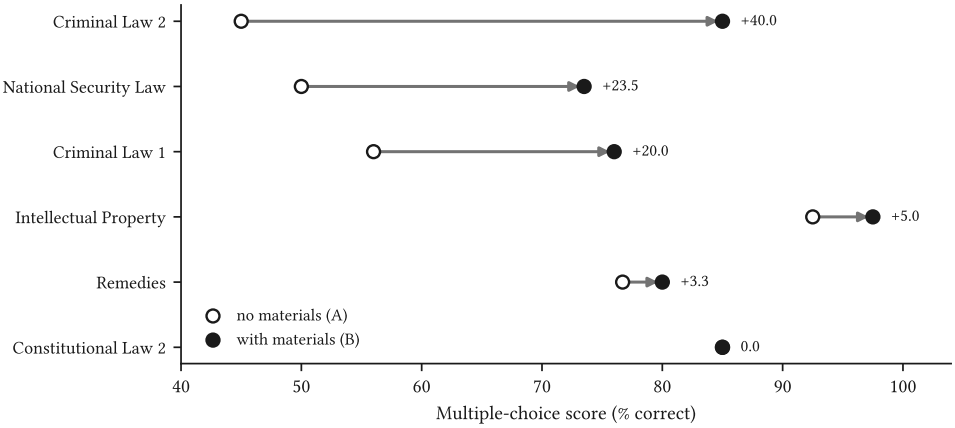

Figure 3: Materials and the multiple-choice score. Each multiple-choice exam with no materials (open) and with materials (filled); the gain runs from nothing to forty points, largest where the no-materials score started lowest.

Now the essays, where the effect nearly vanishes. Across the four essay-only exams with numeric scores, the average materials effect was about four one-hundredths of a standard deviation—statistically and practically indistinguishable from zero. On the Federal Income Tax exam the with-materials arm actually scored a hair *lower*. On the International Law take-home both arms scored identically. On Legislation the materials moved the score a single point out of a hundred. The machine's essay competence, in other words, comes overwhelmingly from training; the outline adds little to prose it can already write.

Then the finding that reframes the whole question. On the National Security Law exam, the materials *helped the multiple-choice score by twenty-three points and hurt the essay score by nineteen.* The with-materials arm wrote a markedly weaker essay than the no-materials arm—the first and sharpest instance in the cohort of materials degrading essay work, plausibly because the dense outline pulled the model toward exhaustive issue-listing and away from the disciplined argument the question rewarded. Averaged into the exam's composite, those two large, opposite effects nearly cancel: the net materials effect on the bottom-line grade was about two-tenths of a standard deviation, a number that conceals everything interesting that happened underneath it.

The lesson generalizes, and it is methodological: the composite is the wrong unit for measuring what materials do. On a blended exam, a near-zero materials effect can mean one of two very different things: the materials did little to either component, or the materials did a great deal to both, in opposite directions. Only component-level analysis distinguishes them. The cohort contains both kinds of "null," and only the component level tells them apart—which is why Figure

4 reports the components, not the composite.

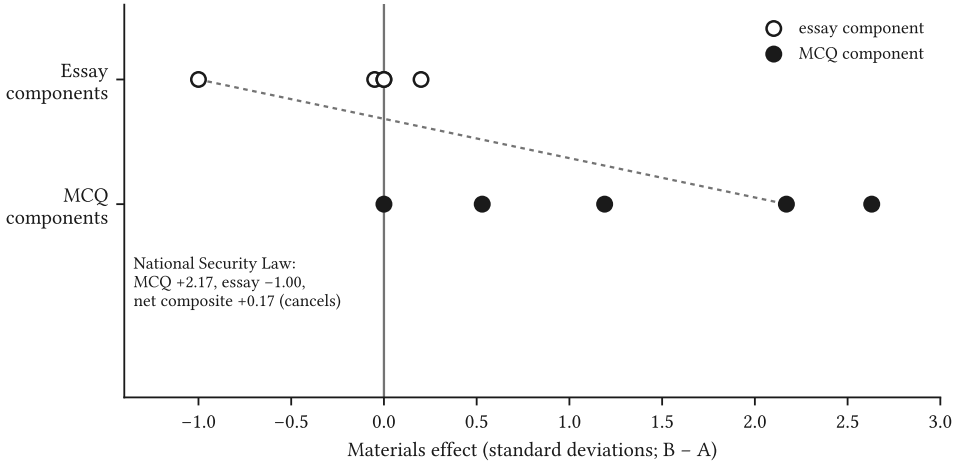

Figure 4: The composite hides the action. Materials barely move the essay components (clustered near zero, one negative) but move the multiple-choice components a great deal. On National Security Law the two run in opposite directions—multiple choice up 2.17 standard deviations, essay down 1.00—and nearly cancel in the averaged composite. Essay components: Federal Income Tax, International Law, the Wire *seminar, Legislation, National Security Law. Multiple-choice components: Constitutional Law 2, Intellectual Property, Criminal Law 1, National Security Law, Criminal Law 2.*

C. Subject matters as much as format

The pre-registered hypothesis framed the materials effect as a function of *format*—multiple-choice versus essay. The fuller picture is that the relevant axis is partly *subject*. The clearest evidence is constitutional law, which the cohort happens to test twice: one essay-only section and one section blending multiple-choice with short answer. On both, the materials effect was small. On the multiple-choice-plus-short- answer section the composite effect was seven-hundredths of a standard deviation, and—crucially—this was not a ceiling artifact: the class scores ranged widely, and the no-materials arm reached the top of the class on the short-answer portion without any course materials at all. Constitutional law is a body of doctrine the model commands fluently from training, so course-specific grounding has little to add. Surveyed after grading, the Constitutional Law 2 grader read the small effect the same way: the materials version "didn't do better," he reasoned, because "my test is geared to publicly available con law and not my idiosyncrasies." The essay-only constitutional-law section is the lone exception to the essay-null pattern—its grader judged the with-materials answer clearly stronger—but he declined to assign numbers, so the judgment is directional rather than quantified, and it cannot enter the pooled estimate.

Put the format and subject observations together and the picture is one of heterogeneity that

resists a single headline. Whether materials help depends on the kind of question (large help on recall-style multiple-choice, little on essays), the subject (little to add where the model is already fluent), and the instrument's headroom (no room to show an effect where both arms already top the class). A random-effects pooling of the materials effect across exams is the honest summary, and it reports exactly this: a positive average effect driven by the multiple-choice components, with substantial cross-exam variance that the format-and-subject story explains better than any single moderator. The machine is helped by an outline the way a strong student is—most where the test rewards recall, least where it rewards the reasoning the student (or model) could already do.

D. The ceiling the machine reaches, and the one it does not

The machine topped curves built to reward coverage, correctness, and polish. But the cohort contains a clean illustration of what those curves miss—on the very exam where the machine's standing was highest.

On the Legislation exam, both AI arms outscored every enrolled student. The top *human* answer, which scored just below them, did things neither machine answer did. The student wrote the exam as a mock judicial opinion, casting the competing schools of statutory interpretation as the feuding judges of a classic jurisprudence hypothetical—down to a judge who, beset by doubt, recuses himself rather than decide.[12] The conceit was an inside joke on material the course had taught, sustained at length and—in the grader's words—delivered with "ridiculously rare creativity and a bit of courage." It was also substantively sharp where the field was thin: the student was among a small minority of the class to resolve a key provision the way the question invited, and nearly alone in naming a particular institutional actor that a careful interpreter should answer to, which the grader called a mark of real "institutional savvy."

None of that is what the machine produced. The AI answers were thorough, fluent, well organized, and—three graders independently observed—uncannily similar to one another. They were not witty, or structurally daring, or original; they did not take the risk of building an argument around a joke and make the joke land. This is the same quality that appears on the detection side as anomalous polish (Part V). The machine reaches the ceiling of the competent, complete, professional answer and stops there. The strongest human work clears a different bar—voice, invention, and judgment about what is worth saying—that the rubric does not score and the model does not, at least yet, supply.

The point is not that the human "won"; on the rubric, the machine scored higher. The point is that the rubric and the machine are calibrated to the same target, and it is not the target that separates a merely excellent legal mind from an exceptional one. Part VI takes up what that implies

[12]The student's anonymity is preserved here, and identifying specifics of the answer have been generalized; the example is used with the grader's account of why it stood out.

for how we ought to assess.

V. Results: Can Faculty Detect the Machine?

The second measured outcome is detection: when a law professor grades her own exam blind, with AI answers mixed into the stack, can she tell? The Wave 1 answer is a qualified yes, and the qualification is the whole story. Faculty detected the machine at a high rate, but the *mechanism* of detection divides into a fixable part and a durable part, and only the durable part tells us anything about whether machine-written legal analysis is identifiable as such.

A. Two detection counts, and why they differ

I counted detection two ways, because the Wave 1 returns forced the distinction. The *reported* count is the detections a grader committed to the record with the return—a flag on the scoresheet, or a structured pre-grading guess. The *detected* count is every instance in which a grader identified an answer as AI during grading at all: the reported cases plus the ones she noticed but did not formally record (often because she understood the protocol to ask her to grade as if the answer were a student's). The detected count contains the reported one—a grader cannot report a machine she did not catch, but she can catch one she does not report.

Across nineteen AI cells graded blind, the reported count was eleven of nineteen and the detected count seventeen of nineteen (Figure 5). (The denominator is cells, not exams: each blind-graded exam contributes its two arms, and the two Criminal Law multiple-choice cells whose AI identity leaked through a build artifact, rather than through blind reading, are excluded from the count.) The two diverge mainly on two exams where the graders spotted the machine immediately but said nothing on the scoresheet because the tell was a formatting artifact they assumed they were meant to ignore. The gap between the two is itself a finding: how often faculty "catch" AI depends heavily on whether you ask them to, and on what they think a detection is for.

Some of what follows—both the detection signals here and the impressions in Part VI—comes from a short, anonymous survey sent to each instructor after grades were in. Four of the ten external graders responded; because they knew the project existed and that AI answers were in their bundle, the survey records single-blind suspicion rather than blinded detection, and I read its answers as illustration, not as a rate.

B. The fixable tells

Now separate the detections by mechanism, because the headline rate is misleading until you do. Of the nine faculty who graded blind, three identified the machine through signals that have nothing to do with the quality of its writing.

The recurring artifact was formatting. The AI pages were single-spaced where the proctoring

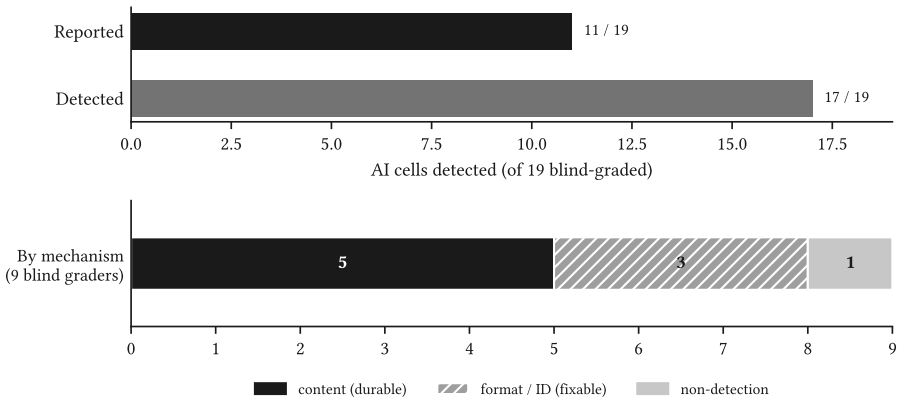

Figure 5: Faculty detection. Top: of the nineteen blind-graded AI cells, the number a grader formally put on the record (reported) and the larger number caught during grading whether or not it was written down (detected, which contains the reported cases). Bottom: the nine blind graders by what gave the machine away—the writing itself (durable) versus formatting and identification artifacts that any competent build will fix.

software double-spaces student exports; the body font differed; the per-question page annotations the software adds to student work were missing. One grader was explicit: "I spotted both AI exams as AI exams but only because of the formatting . . . but thought I was supposed to ignore that." Another caught them on a different font and single spacing, plus an identification-number tell: the reserved examination numbers were "the last two exam numbers," sitting conspicuously at the end of her ID-sorted sheet. A third grader's certainty was confirmed, in part, by a reversed-glyph artifact—the model's answer "consistently put opening quotation marks backwards." None of these is a property of legal analysis. Each is a bug in how the answers were built and numbered, and each is fixed for Wave 2 (Part VIII). To the extent the Wave 1 detection rate rests on them, it overstates how identifiable machine-written legal writing actually is. The survey underscored how shallow these tells run: the *Wire* seminar's grader, who had braced for "lots of bullet points and bureaucratese" and instead found the answers "well written, concrete, tight, convincing," reported that the only thing needed to hide them was to "format them like the real exams . . . to conceal authorship."

A related leak deserves its own mention because it is both a tell and a contamination risk. On two exams the with-materials answer carried a vestige of its own construction into the delivered text: in one, a one-line "Course materials are attached" preamble that survived the audit because it was fused to the first sentence; in another, the model broke character mid-answer to write that it "might cite this line of cases, except I'm not sure what [the professor] would think"—naming the grader. These are not stylistic signals; they are construction artifacts, and they identify the with-materials arm specifically. They are fixed the same way the formatting tells are.

C. The durable signal

What remains, after stripping the artifacts, is the detection that should make us think. Five of the nine graders identified the machine on the content of its writing, and their accounts converge on two cues.

The first and strongest is *inter-arm similarity.* The two AI answers to a given exam—generated in complete isolation, by separate model invocations that never saw each other—frequently read as though one author wrote both. Three different graders, independently, used almost the same words. One: the two exams "have similar textual flows," so much so that "if it hadn't been AI I would have accused those exams of cheating." Another: "very similar to each other and quite different from what real students produced under time pressure." A third: "the tone and voice . . . seemed unusually similar to me; they could have been authored by the same individual or entity." This is a genuine signature of machine authorship, and it is not easily engineered away, because it is a property of the model's stable voice rather than of any fixable build step. Two students do not write alike; two runs of the same model do.

The second content cue is *register.* Time-pressured, closed-book student writing has a texture—uneven, with spelling slips and abandoned sentences—that the machine's prose does not. One grader flagged the AI precisely because it was "incredibly thorough and well written" in a setting where even strong real answers carried "grammatical and spelling errors." Polish, in the exam room, is itself anomalous.

The most striking detection in the cohort combined both cues and ended in a decision not to grade. The essay-only constitutional-law grader identified both AI answers early in his reading, named which arm had received materials and which had not (the only grader to get the condition right, not just the identity), and then declined to grade them. His concern was that once he could see which answers were machine-written, scoring them on the same curve as his students would no longer be an unbiased exercise—his judgment compromised by knowing what he was reading. His detection rested on the writing and analysis; the formatting and numbering tells, he said, only confirmed what was already apparent. That is the durable signal operating at full strength—a reader who knows the subject cold, identifying the machine from the shape of its thought, and finding the identification disqualifying enough to set the scoring aside.

And yet one contrary data point keeps the durable signal honest. On the International Law take-home, the grader caught nothing. He scored the printed exams blind, in pencil, and identified the AI answers only when he went to enter the scores and found the registrar's spreadsheet had no rows for the reserved numbers—a logistics accident, not a detection. "I swear I was just transcribing at that point." A capable grader, given a clean take-home and no procedural tell, did not flag machine-written work he had just placed in the top quartile. Detection is high in this cohort, but it is not automatic, and the take-home format—more time, more polish expected from students, no proctoring artifacts—is where it failed.

VI. Discussion: What Are We Measuring?

The question Parts IV and V force is not first of all what to do, but what we are now measuring, and it is the one I want to put to colleagues: if a machine can ace the exam, is the exam still measuring the thing we mean it to measure? I offer this as a problem to sit with, not a verdict, and not a program for reform. The more urgent ask is not a particular redesign but intentionality—that each of us be clear about what a given assessment is meant to measure, and honest that the rise of AI is widening an already-real gap between the law-school exam and the work of a practicing lawyer. What follows from that will, and should, vary by course and by colleague. For some it may argue for rethinking assessment top to bottom; for others, for narrower adjustments; for others still, only for naming the mismatch plainly and grading with it in view. Reasonable faculty, looking at the same evidence, will land in different places. Any redesign belongs to the person who owns the course; what a study like this can contribute is evidence about where the old assumptions have quietly stopped holding.

A. What the exam has stopped measuring

Begin with the finding that should unsettle us most. On several exams the *no-materials* arm—the model working from training alone, with no syllabus, no outline, no knowledge of the course—reached the top of the class, and the materials I supplied barely moved its essays. Put those two facts together and the implication is uncomfortable: when a model that has never seen your course writes a top-of-the-class answer to your final, the exam has stopped measuring mastery of *your course* and started measuring fluency in a *genre*—the time-limited, word-limited issue-spotter, the doctrinal multiple-choice block—that the machine has thoroughly mastered. A high-stakes instrument whose job is to discriminate this cohort's command of this material is, on this evidence, discriminating something else. And it discriminates poorly even at that: because the format rewards organized, doctrinally complete exposition—the exact register the machine maxes—the ceiling is now machine-reachable, and the strongest students pile up against it alongside the machine, where the instrument can no longer tell them apart.

There is a deflationary reading of all this, and it deserves a hearing because it is partly right. The law-school exam was never a faithful model of what lawyers do. It is a contrivance built for measurement—time-limited, closed-universe, written alone against the clock—precisely because those constraints make grading tractable and comparable, not because practice looks anything like that. On that view the machine's success is less alarming than the headline sounds: it has mastered an artificial genre, not the open-ended, collaborative, judgment-laden, client-facing work of an actual lawyer, and we should resist reading "aces the exam" as "can do the job." I think that reading is right as far as it goes. But it cuts toward the worry rather than away from it. If the instrument was already a loose proxy for the capacities we care about, and a machine can now max the proxy without the underlying capacity, then the case for treating that instrument as a high-stakes discriminator of *human* ability is weaker still—not because AI has arrived, but because

AI has exposed how much the exam was already asking us to assume. And the gap only widens from here: as practicing lawyers increasingly work *with* these tools, the exam's defining artifice—the unaided human, alone against the clock—represents the job less faithfully each year, even as we lean on it to sort the people headed into that job.

A further objection runs the other way—the calculator. We still teach and test arithmetic though every phone can do it, because fluency in the foundation is what lets a person do the higher-order work and catch the machine's errors. The objection is right—it is the one a skeptical colleague raises first—and it deserves a straight answer. And yet it does not defend the *high-stakes, end-of-term, one-shot* exam; it defends *teaching and checking the foundation.* Those are different jobs, and the traditional final conflates them. One way to pull them apart—and the direction several faculty in this cohort are already moving—treats foundational competence (issue-spotting, doctrinal recall, the organized exposition the machine now supplies on demand) as worth building but as a low bar a machine clears: confirming that a student has cleared it need not take a three-hour proctored finale, and can live in formative, low-stakes checks across the term, where the point is to build the foundation rather than to rank the class on it. That would free the high-stakes instrument for the job the format now does badly: measuring judgment, synthesis, and the things a lawyer is actually paid for. I raise it as the natural drift of the measurement problem, not as a reform the data alone can carry.

B. The moving target

If faculty do redesign—and much of this decoupling is not news to them, who have felt the ground move and are already at it—the target will not hold still. The cohort holds a clear example: on one exam the instructor built the central problem around invented law with no analog in any training corpus, and it is the one place in the study where the machine sat near the bottom of the class on the applied question while still topping the doctrinal one. The design resisted retrieval, and the data rewarded it. And yet instinct is an unreliable guide here, which is precisely why an empirical study matters. The intuitive moves toward "AI-proof" assessment often backfire. Opening the question up, making it more reflective, pushing it toward theory—these feel like they should defeat a machine, and they are exactly where the machine is strongest: the wide-open reflective essays and the normative questions in this cohort were where both arms most often reached the ceiling. One grader's first reaction—that "wide-open essay questions are just too easy for AI to hack"—is the lesson in miniature. What actually bit the machine was narrow and specific: genuine novelty that could not be retrieved. A redesign built on instinct alone just sets up the next exam the machine aces.

Which raises the hardest problem, and the one I flag rather than pretend to solve: the capability is advancing extremely fast. A strategy of "find what the model is bad at this year and test that" is a treadmill—this year's retrieval-resistant problem is next year's solved one, and I would ex-

pect the very gap the fictional-statute exam exploited to narrow in the next wave of this study.[13] So for faculty who do take up the redesign, the durable principles cannot be a snapshot of current weaknesses. Three hold up better than instinct. The first is assessing what we intrinsically value and what practice requires—judgment, ethical reasoning, the ability to interrogate a problem in real time, and above all the capacity to *direct and correct* a machine rather than be replaced by it—skills worth certifying in the human even as the model improves at them. The second is assessing the *human-plus-AI system*, not the unaided human, because that is the unit of real practice: an "open-AI" assessment that hands the student the model's answer and grades what the student *adds*—the errors caught (including the fabricated citations the model still produces), the comprehensive-but-unfocused draft sharpened into an argument—tests the supervisory judgment the calculator objection says we still need, and it turns the machine's own blind spot into the task, since recognizing what is wrong with a fluent answer is exactly what the model itself cannot do. The third is favoring process over one-shot product: as the product becomes machine-replicable, the cultivated process—drafts, revision, a position defended aloud—is what we are after. And because the target moves, assessment design has to be revisited as the capability does. This study is built as successive waves for that reason; the posture it recommends is less a fixed redesign than a habit of measuring as the ground shifts.

C. What the graders said

The post-grading survey put this Part's themes in the faculty's own voice. The Constitutional Law 2 grader had set out, he wrote, "to write an exam that would sort students fairly in an environment in which they couldn't use AI, not an exam that would be hard for AI," and was "surprised by their theoretical sophistication": the machine's "theory answers were much better than most students'." He read that capability as an opportunity as much as a threat—a model that strong "could be a very effective interactive tutor," one that might "nudge student exam preparation in the direction of learning constitutional law rather than learning how to take my exam." The International Law grader put the policing point flatly: "there is no way to 'AI proof' an exam," and the take-home formats that try are the least defensible of all—"back to in-class exams." And the Remedies grader named the bind that makes redesign both urgent and uncomfortable: squeezed between disability accommodations and AI, he now leans on closed-universe, time-pressured exams that, he conceded, put students "behind AI structurally," and he ended where this Part does—"we have to find a way to give assessments that students can beat." Four returns are not a finding, but they corroborate the argument from the inside: the instrument may be measuring something other than

[13] The pace, and even the reality, of "reasoning" gains is contested. Compare the rapid model-over-model improvement in the Stanford study's ranking, *supra* note 7, with Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio & Mehrdad Farajtabar, *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*, arXiv:2506.06941 (2025) (arguing reasoning-model accuracy collapses past a complexity threshold). The design implication does not turn on resolving that debate; it needs only that the capability is moving fast enough to make a snapshot of current weaknesses a poor foundation for assessment design.

what we think, and the instinctive fixes misfire. Two of the four reach for the same fallback—lock the exam down, go back to in-class—which raises the last question: if you cannot out-design the machine, can you at least police it?

D. Design, not security

The honest answer is no, or not for long. Detection is real but contingent: the durable signal is available only to an expert reading multiple machine answers, it does not survive the single submission a real cheating student would file, and the automated detectors the integrity literature distrusts are worse still. Locking exams down harder is an arms race the take-home result suggests the machine quietly wins, and it does nothing about whether the locked-down exam measures what its author intends. Whatever the answer turns out to be, it lies in being intentional about the design of the assessment, not in the security around it.

So what is worth protecting? Two studies this season, from different schools and two directions—preference in the tutoring channel, performance on the graded exam—arrive at the same place: a frontier model now produces answers that competent faculty grade at or above the class median and often prefer. That is a prompt, not a crisis. The honest answer to the title is yes, the machine can ace your exam, often, and largely on the strength of what it already knows. And yet on one exam the answer its grader singled out as exceptional was a student's—a mock judicial opinion built around a creative conceit and sustained with wit and nerve, the kind of answer the rubric does not reward and the model does not produce. I do not claim that gap is permanent. I claim that today our instruments are calibrated to a ceiling the machine has reached, and that what the exam can no longer cleanly discriminate is exactly the human work that lives above it. The question for legal educators is no longer whether the machine can do what we test, but whether the exam is still measuring what we have long assumed it measures.

VII. Limitations

This is a pilot, and its limitations are real. I catalog them rather than minimize them, both because the results should be read in their light and because each one defines a piece of the Wave 2 design.

Start with scope. Every result here is Claude Opus 4.7, accessed as a single model from a single vendor, so nothing generalizes to other systems without testing them—Wave 2 adds that cross-vendor coverage. The one vendor is, on an independent benchmark, among the strongest currently ranked: an LLM-as-judge extension of the recent Stanford study, its judge first validated against the human-majority preference, ranked Claude Opus 4.7 first among roughly a dozen systems,[14] so these results are unlikely to understate frontier capability, though that benchmark measures

[14] Salinas, *supra* note 7. In the study's LLM-as-judge extension — its judge first validated against the human-majority preference — Claude Opus 4.7 ranked first among roughly a dozen systems. The study's separate expert-validated *human* evaluation compared only two systems (Gemini 2.5 Pro and NotebookLM).

a different task than graded exam-writing. Isolation between conditions also required a single generation per cell, so run-to-run variance is unestimable from this data, and a different sample from the same model might land a few points elsewhere; Wave 2 adds replication. And the cohort is small—eleven exams, one term, one school, faculty who volunteered—so the pooled estimates carry wide intervals, and the descriptive pattern, not any single significance test, is the contribution.

The treatment was not uniform. The with-materials arm was "whatever course materials could be assembled," which differed across exams—sometimes a prior student's outline, sometimes a richer AI-synthesized one. That muddies the cross-exam materials comparison, and a minimum-materials standard is a principal Wave 2 target.

The blinding was imperfect, and in a way that bears directly on the detection results. The formatting, font, spacing, and identification-number tells of Part V.B mean several Wave 1 detections rode on artifacts rather than prose, so the detection rate overstates prose-level detectability—and the reserved examination numbers, sitting just above and clustered within the real student range, were a non-prose tell that drove at least one detection outright (Wave 2 intersperses them). The departures from clean blinding are flagged wherever they bear on a result: one exam was graded by an instructor who knew during grading that the cells were AI, though not which arm had materials, so the materials contrast survives; one instructor declined to assign numbers; and one exam was graded by the author, at arm's length, with the conflict noted. The pre-registered sensitivity analyses re-run the headline estimate dropping the affected exams. The post-grading grader survey carries its own caveat: only four of the ten external graders responded, a self-selected minority, and its single-blind framing captures suspicion rather than blinded detection—so its quotes illustrate the cohort's themes but cannot stand as a rate.

A final limitation is about what the study does not measure. It asks whether the machine can produce graded exam answers; it says nothing about whether students who use these tools learn more or less. Those are different questions, and the emerging evidence cuts against the optimistic reading: heavy reliance on model-generated summaries appears to produce shallower understanding than traditional research does.[15] A machine that can ace the exam is not evidence that the student who leans on it has learned the subject.

VIII. The Next Wave (Fall 2026)

Wave 1's second purpose was to expose what a confirmatory study should tighten, and it did; the lessons-learned record converts directly into the Wave 2 design.

The first job is to close the blinding leaks, since they are the cheapest thing standing between this design and a clean detection measurement. The build script will match the proctor-

[15] See Shiri Melumad & Jin Ho Yun, *Experimental Evidence of the Effects of Large Language Models Versus Web Search on Depth of Learning*, 4 PNAS Nexus pgaf316 (2025).

ing software's double-spacing, font, and per-question page annotations; the reserved examination numbers will be interspersed within the real student range rather than stacked above it, and pre-inserted into the grading roster so no spreadsheet gap unmasks them; and the construction artifacts—the "materials attached" preamble, the reversed quotation glyph, the in-answer references to the grader—will be stripped. Each removes a fixable detection vector, leaving Wave 2 to measure the durable, content-based signal cleanly.

The design itself widens on several fronts. Cross-vendor coverage tests whether the findings are Claude-specific, and multiple runs per cell put error bars on the within-cell variance Wave 1 cannot estimate. The treatment gets standardized as well: a minimum-materials standard—ideally with a controlled comparison of student-prepared versus AI-synthesized outlines on the same exam—replaces Wave 1's opportunistic assembly, so the materials contrast means the same thing across exams. And the materials hypothesis will be pre-registered where it actually lives: Wave 1's central analytic lesson is that the effect hides in the components and cancels in the composite, so Wave 2 registers the hypothesis at the *component* level—multiple-choice effect versus essay effect—rather than on blended composites.

Finally, detection becomes a measured task rather than something reconstructed from emails. The pre-grading guess channel worked—faculty find it intuitive, and three graders volunteered correct guesses unprompted—so Wave 2 formalizes it: a structured flag-and-rationale paired with a pre-grading guess, so detection accuracy and mechanism are measured directly across the full cohort.

Wave 2 runs in Fall 2026. Its job is to take the three Wave 1 findings—competitive performance, a component-located materials effect, and a content-based detection signal beneath a layer of fixable artifacts—and test whether they hold when the artifacts are gone, the model is not the only one in the room, and every cell is run more than once.

Conclusion

A frontier language model sat for eleven real Penn Carey Law final exams, graded blind on the curve, and performed like a strong student—passing everywhere, at or above the median everywhere, in the top decile most of the time, and above every enrolled student on the two exams scored on an open, non-saturated scale. It did this largely on training rather than on the course materials I supplied; the materials helped where exams reward recall and barely at all where they reward reasoning, and on one exam they helped one half of the exam while hurting the other. Faculty mostly caught the machine, but mostly through formatting and numbering artifacts that any competent build will remove; the detection that survives is the machine's stable voice and unhuman polish, visible to an expert reader and, in one case, disqualifying enough that the grader set the scoring aside.

None of this settles the normative question it raises—whether the exam still measures what we have long taken it to measure once a machine can ace it blind, and how much it measured even before. But it answers the empirical one the project asked. Can AI ace your exam? On the evidence of Spring 2026 at one law school: yes, and more easily than the "can it pass" framing suggested. That a machine can max an instrument built for a narrower purpose says as much about the limits of that instrument as about the machine. The harder work—being intentional about what each instrument is meant to measure, and honest about a mismatch AI only widens—is where reasonable colleagues will differ, and where Wave 2 begins in the fall.

Appendix A. Per-exam results

The table below is the cohort summary; the canonical source is the project's `scoring-results.csv`. Faculty are anonymized by subject in this draft (see the project name registry). Materials effect is the standardized difference (with-materials z minus no-materials z). One exam (Criminal Law, four-component) returned only its multiple-choice result; its essays were not graded, so it contributes no composite.

Exam (subject)	Format	Real n	No-materials	With-materials	Materials effect	Detection (mechanism)
Federal Income Tax	essay-only	37	159/200, z +1.76, rank 2	158/200, z +1.71, rank 3	-0.05 z	yes (content: shared blindspot)
Remedies	MCQ+essay (+C)	41	3.4 GPA, z +0.34, rank 18	3.799 GPA, z +1.65, rank 4	+1.31 z	yes (content: out-of-syllabus + inter-arm)
Intellectual Property	MCQ+essay	77	83.30/100, z +2.19, rank 1	91.30/100, z +3.12, rank 1	+0.93 z	n/a (operator-graded)
National Security Law	MCQ+essay	36	121.0/200, z +0.00	125.53/200, z +0.17	+0.17 z (MCQ +2.17 / essay -1.00)	yes (format + ID)
International Law	essay-only (take-home)	52	51, z +0.79, rank 9	51, z +0.79, rank 9	0.00 z	no (clean non-detection)
Legislation	essay-only	84	99/100, z +2.47	100/100, z +2.67	+0.20 z	yes (content pre-guess)
Constitutional Law 1	essay-only	92	(declined to grade)	(declined; "best of the lot")	qual. B>A	yes (content; correct on arm identity)
Constitutional Law 2	MCQ+essay	87	105/120, z +1.65, rank 5	106/120, z +1.72, rank 2	+0.07 z	yes (content + pre-guess)
Criminal Justice (The Wire)	essay-only	12	40/40, z +1.85	40/40, z +1.85	0.00 z (ceiling)	detected (format); 0/2 reported
Criminal Law 1	MCQ+essay	90	composite z +0.02, rank 39	composite z +1.10, rank 19	+1.08 z	detected (format); 0/2 reported

Exam (subject)	Format	Real n	No-materials	With-materials	Materials effect	Detection (mechanism)
Criminal Law 2	MCQ+essay	85 (MCQ)	MCQ 9/20, z -0.26	MCQ 17/20, z +2.37	MCQ +2.63 z; essays not graded	artifact-driven leak (excluded)

Across the six exams with a multiple-choice component, materials raised the score by an average of +14 to +15 points, with a range of 0 to +40; the two largest were +40 (a Criminal Law exam, from the class mean to the class ceiling) and +23.5 (National Security Law). On detection, the totals were 11 of 19 reported and 17 of 19 detected, with the mechanism splitting 5 content, 3 format-tell, and 1 clean non-detection.

Appendix B. The essay prompt template

Every essay cell was generated from the template below, reproduced verbatim. Only the bracketed slots (`[COURSE NAME]`, `[PASTE THE FULL EXAM QUESTION HERE]`) and the attached materials varied across cells; in the no-materials arm, no documents accompanied the prompt and the model was instructed not to signal their absence. Multiple-choice cells used a parallel template that produced a letter-only answer set and omitted the prose-format and IRAC instructions.

```
<role>
You are a strong law student sitting for the final exam in [COURSE NAME]. You have done the
     reading, attended class, and built an outline. You are writing under time pressure.
</role>

<task>
Write a complete answer to the exam question in <exam_question>. Your answer will be graded
     blind, on the same curve as real student answers, by the professor who wrote the
    question. Aim for solid A-/A range performance.
</task>

<course_materials>
Course materials may be attached (syllabus, readings, slides, handouts, prior exams). Begin
     by noting whether course materials are attached.

If materials are attached, treat them as your primary source. They will not always be the
    complete set of materials for the course.

If no materials are attached, proceed using widely-recognized authority. Do not signal this
     absence in your answer; simply rely on rules and cases you are confident are real and
    well-established.
</course_materials>

<success_criteria>
```

```
Your answer should (a) read as written by a real student under exam conditions, and (b)
    earn a strong grade on the curve.

A strong answer has three qualities. Issue depth: it identifies the subtle issues most
    students miss. Rule precision: it states rules in specific terms, not vague paraphrases
    . Fact application: it ties each rule to the specific facts in the question, repeatedly
    .

Pick one issue to analyze more sharply than the others, with an observation that goes
    beyond the standard treatment. Do not give every issue identical depth.
</success_criteria>

<format>
Follow every instruction in the exam question, including word limits, format requirements,
    and the precise call of the question. If the question asks for the strongest argument
    for one side, do not write a balanced analysis. If it asks for advice to a client, do
    not write a judicial opinion.

If the question specifies a word limit, treat it as a ceiling. Aim for 85-95% of the limit.
     Do not pad to reach the limit. If you exceed it, cut from the weakest issue, not the
    strongest.

Write in continuous prose. No headers, sub-headers, or bullet points unless the question
    requires them. Skip the introduction and conclusion. Begin with the first substantive
    sentence of analysis.

Use compressed IRAC. After first full reference to a case, use short form (Palsgraf, not
    Palsgraf v. Long Island R.R.). Do not italicize case names. Do not use Bluebook
    formatting.
</format>

<style>
Use abbreviations a student would naturally use in this subject area, drawn from the course
     materials where attached. Do not force generic abbreviations.

Vary sentence length. Use contractions occasionally. Use casual exam transitions ("Here," "
    On these facts," "The closer question is," "D's best argument is").

Do not use em dashes. Do not use the following stock phrases: "it is worth noting," "
    notably," "importantly," "furthermore," "moreover," "in conclusion," "this raises
    important questions," "there are strong arguments on both sides," "this is a nuanced
    issue," "underscores," "underpins," "pivotal," "multifaceted," "comprehensive," "robust
    ." If you find yourself reaching for one of these, rewrite the sentence.

Take positions on close calls. Commit. "D's best argument is X but it likely fails because
    Y" beats "reasonable minds could differ." Do not hedge at the conclusion of an issue.

Where useful, address both sides of an ambiguous issue briefly, then commit to one. Do not
```

```
    present a balanced analysis without committing.
</style>

<citation_honesty>
Do not fabricate case names, statutes, or citations. This is the single most important rule
    .

Where course materials are attached, cite cases and rules from those materials. Prefer
    rules and framings that appear in the attached materials over those drawn from general
    legal knowledge, even when the latter would be more elegant. Match the vocabulary of
    the attached materials where possible.

Where materials are not attached, use widely-recognized authority you are confident is real
    .

If you remember a rule but are not certain of the source, do not invent one. Write "the
    rule in this course is," "per the casebook," or "class discussion established that."
    Real students who blank on a case name do this routinely and are not penalized.

If you cite a case, the case must exist and the rule must be substantially what that case
    actually held. When in doubt, describe the rule without naming the case.
</citation_honesty>

<execute>
Output only the exam answer itself. No preamble, no meta-commentary, no headers.
</execute>

<exam_question>
[PASTE THE FULL EXAM QUESTION HERE]
</exam_question>
```

Appendix C. The pre-registered analysis plan (operative excerpt)

The plan below was locked on May 7, 2026—after all cells were generated and packaged, and before the bulk of the grades returned (three of the eventual numeric returns were in hand). It is reproduced here in operative form: the decisions fixed in advance, stripped of the surrounding methodology narrative (which appears in Part III). Any change after the lock date is recorded, with its rationale, in the project's decision log; the pre-registered version remains in the project's version history as the audit trail.

Primary contrast. No-materials (A) versus with-materials (B), tested within the AI submissions. Each exam is one independent observation; per-component cells contribute to the arm total via the course's own syllabus weighting. The single per-exam C arm is a within-exam secondary contrast, not pooled. The AI-versus-real-student comparison is descriptive context, not the contrast under

test.

Outcome variables, in priority order: (1) per-arm raw score on the exam's native scale; (2) per-arm z-score against the real-student-only distribution—`(AI - real mean) / real SD`, sample SD with an n-1 divisor, the AI submissions excluded from the comparator—which is the primary cross-exam unit; (3) per-arm percentile and rank against the real-student-only distribution; (4) the faculty AI-flag (binary, per cell).

Materials effect, reported three ways per exam: raw (`B - A`), standardized (`B_z - A_z`, the primary effect-size unit), and rank.

Format stratification and its hypothesis. The cohort splits into an essay-only stratum and an MCQ-bearing stratum; for MCQ-bearing exams the multiple-choice-component, essay-component, and composite effects are reported separately. The pre-registered hypothesis: *the materials effect is larger in MCQ-bearing exams than in essay-only exams.*

Pooling. Random-effects meta-analysis across exams on the standardized effect, with heterogeneity reported (I^2, Q-test). If I^2 exceeds 50%, the pooled estimate is read descriptively and the format-stratified estimates become the principal report. Pre-registered moderators: materials count and type, delivery mode (in-class versus take-home), and the faculty AI-flag.

Multiple comparisons. The pre-registered test family comprised the per-exam A-versus-B contrasts, the cohort pooled estimate, the two stratified estimates, and the four moderator analyses, with Holm-Bonferroni correction. Descriptive effect sizes with 95% confidence intervals are primary; formal significance tests are supplementary, and the headline claim does not depend on any single test crossing a threshold.

Three locked definitions of "the AI passed":

- *Competitive*—both arms earn passing letter grades on at least 80% of exams.
- *At or above median*—one arm reaches the 50th percentile or better on at least 50% of exams.
- *Wins*—one arm reaches the top decile on at least 50% of exams.

The paper reports all three; which is foregrounded depends on the data.

Pre-specified sensitivity analyses (reported in supplementary materials): drop the per-exam C arm; restrict to the standard two-document materials condition; drop the author-graded exam; restrict to exams delivered after the packaging-format fix; and restrict to closed-book exams, where the model's information set is most directly comparable to the students'.