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Not for publication or circulation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`

Extended Abstract

Ask whether an artificial intelligence can pass a law school exam and you have asked a question the technology answered some time ago. The sharper question is the one this study put to the test: where does a machine land when its work is graded blind, on the curve, against the students who actually sat for the exam? Over the Spring 2026 examination period at the University of Pennsylvania Carey Law School, a frontier large language model sat for eleven real final exams — Federal Income Tax, Remedies, two sections each of Constitutional Law and Criminal Law, Intellectual Property, Legislation, National Security Law, International Law, and a seminar on criminal justice and *The Wire*. I packaged the machine's answers to look like ordinary student submissions, gave them reserved examination numbers the registrar guaranteed would not collide with any real student, and dropped them into the live grading stack; the faculty who wrote each exam graded blind, on their own rubrics, and reported back. Each exam was answered twice under otherwise identical conditions: once with no course materials in the model's context, on its training alone, and once with the syllabus and an outline supplied. The design is the point. 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, so this study removed the flattery: unreleased exams, blind expert grading, the real class curve as the comparator (recomputed with the AI answers excluded), and an analysis plan pre-registered before the grades came back. One model, one vendor — Claude Opus 4.7 — by deliberate choice. This is the first of several planned waves.

[†]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.

Three findings organize what came back. First, the machine is competitive at the top, not merely passing. 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 — and on two, both scored on an open, non-saturated scale, it outscored every enrolled student. On Legislation it scored 99 and 100 out of 100, above all eighty-four students in the class; on Intellectual Property it took the top composite in a class of seventy-seven, in both arms. One Constitutional Law grader called the machine's answer "the best of the lot" in a field of ninety-two. The profile that recurs, in the words of the International Law grader, is a familiar one: strong on coverage and issue-spotting, weaker on the subtle argument and the connection drawn across issues.

Second, the value of course materials is real but concentrated, and the bottom-line grade hides it. On the multiple-choice components the materials help enormously: an average of fourteen to fifteen points, and on the hardest instrument a forty-point swing, from the class average to the class ceiling, better than two and a half standard deviations. On the essays they do almost nothing — the average effect across the essay-only exams is about four one-hundredths of a standard deviation, because the machine's essay competence comes from training, not from the outline. And on one exam the two effects ran in opposite directions: materials helped the multiple-choice score by twenty-three points and hurt the essay by nineteen, so they nearly cancel in the averaged composite. The lesson generalizes, and it is methodological: the composite score is the wrong unit for measuring what materials do, because a near-zero net can mean either small effects on both halves or large offsetting ones. The subject matters as much as format. In constitutional law, a body of doctrine the model apparently commands fluently from training, the no-materials arm reached the top of the class with no course grounding at all.

Third, faculty can tell — but mostly for reasons that will not last. The recorded detection rate was eleven of nineteen blind cells, the internal rate seventeen of nineteen, and the gap between them is itself a finding. Separate the detections by mechanism and most of the early ones ride on artifacts that have nothing to do with the writing: single spacing where the proctoring software double-spaces, a different font, reserved ID numbers sitting conspicuously at the end of a sorted list, in one case opening quotation marks rendered backwards. Every one of these is a build-script bug, fixable before the next wave. What survives the fix is harder to engineer away and more interesting: the two independently generated answers to a single exam often read as though one hand wrote both — three graders, separately, called the pair the same "individual or entity" — and the prose carries a thoroughness and polish that time-pressured student work, with its spelling slips and ragged edges, does not. One grader identified both AI answers, named correctly which arm had received the materials, and then declined to grade them, concerned that scoring work they could see was machine-written would no longer be an unbiased exercise.

A quieter thread runs underneath all three. The machine tops curves built to reward coverage and polish, but on the Legislation exam the grader singled out one human answer that did what neither machine answer did: a mock judicial opinion that cast the schools of statutory interpre-

tation as the feuding judges of a classic jurisprudence hypothetical, sustained at length with wit and nerve. It is one observation, from one course and one reader, and I do not lean on it for more than it is worth. And yet it points past what the rubric scores. The machine reaches the ceiling of the competent, comprehensive, professional answer and stops there; the heights above it — voice, invention, judgment about what is worth saying — it did not, here, reach. A recent Stanford-led study, released this spring, finds the same from the tutoring channel: grading blind, law professors prefer AI answers to their colleagues'.

The empirical answer, then, is yes. The harder question is not what to do but what the exam is now measuring, and it is the one I want to put to colleagues: if a machine that never took your course can write a top-of-the-class answer to your final, it is no longer clear the final is measuring mastery of your course rather than fluency in a genre — the time-limited issue-spotter and the doctrinal multiple-choice block — that the machine has already thoroughly mastered. There is a deflationary reading worth taking seriously: the law-school exam was never a faithful model of practice but a contrivance built for grading, so a machine acing it tells us less about lawyering than the headline implies. That reading is right as far as it goes, and it cuts toward the worry rather than away from it — a loose proxy a machine can now max without the underlying capacity has a weaker claim to its high-stakes role, not a stronger one, and the gap only widens as practicing lawyers increasingly work with these tools. The more urgent ask is not a particular fix but intentionality: that faculty be clear about what each assessment is meant to measure, and honest about the widening mismatch. What follows will, and should, vary — wholesale redesign for some courses, narrower adjustment for others, for others still only naming the gap and grading with it in view — and reasonable colleagues will differ. The obvious objection is the calculator: we still teach and test arithmetic though every phone can do it. The objection is right, and it defends teaching and assessing the foundational skills, not the four-hour in-class final; those are different jobs, and the traditional exam conflates them. One response, and the one several faculty in this cohort are already taking, decouples them: foundational work moves into formative, low-stakes checks across the term, freeing the high-stakes instrument for the judgment and synthesis a lawyer must also have. But instinct is an unreliable guide to that redesign, and the data show why: some intuitive moves toward "AI-proof" assessment — opening the question up, making it reflective, pushing it toward theory — are exactly where the machine is strongest. So if the redesign comes, its durable principles are not a snapshot of this year's model's weaknesses but an approach: assessing what we intrinsically value, assessing the human-plus-AI system rather than the unaided human, and favoring process over one-shot product.

This is a pilot, and I treat its limits as real: one model and one vendor, a single run per cell, materials that varied across exams, eleven volunteered exams, and a blinding that the formatting and numbering tells left imperfect. Each helps define a piece of the second wave (Fall 2026), which aims to close the leaks, adds models and repeated runs, standardizes the materials, and turns detection into a measured task. So: can AI ace your exam? On the evidence of one law school in the

spring of 2026, yes — often, and largely on the strength of what it already knew. 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.