Can AI Ace Your Exam?

A Penn Law Experiment

R. Polk Wagner

Early Draft · June 2026

Penn Carey Law
University of Pennsylvania

Special Thanks

- Tulio Tagliaferri (Penn Carey Law J.D. 2027), for his substantial contributions to this project
- Claire Wallace and the Academic Affairs team
- The ten faculty who volunteered their exams for this study (and agreed to grade extra exams)

What AI Already Aces

- The bar exam, near the top
- The LSAT and admissions tests
- *Public, standardized benchmarks*
- *All built to be machine-scored*

What We Know Far Less About

- A real (elite) law school final exam
- Multi-format, written for one class, never released
- Graded blind by its author
- Ranked against enrolled students

Acing a public, standardized test is well-established. Acing a real exam a professor wrote, on the same curve as students, is not.

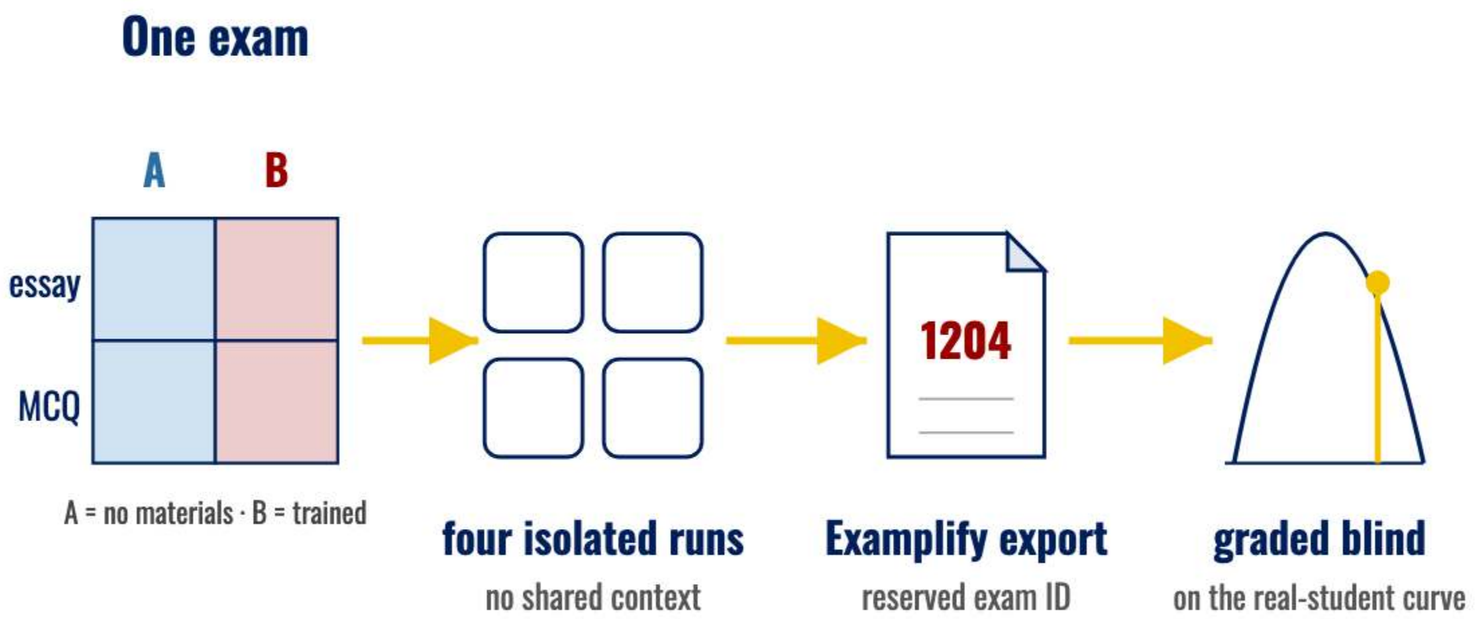
How does a current AI model do on actual law school exams, graded blind on the same rubric and distribution as the students?

How it was done

The Experiment

- Eleven real Penn Carey Law final exams, Spring 2026
- Each component, essay or MCQ, answered twice: no course materials vs. trained (syllabus and an outline)
- One frontier model, one isolated run per cell:
  - Claude Opus 4.7 (1M-token context)
  - knowledge cutoff January 2026; default sampling
  - run via the Claude Code CLI
- Essays graded blind by the faculty who wrote the exam, on the real-student curve

From exam to graded score



Four isolated cells per exam, with no cross-visibility.

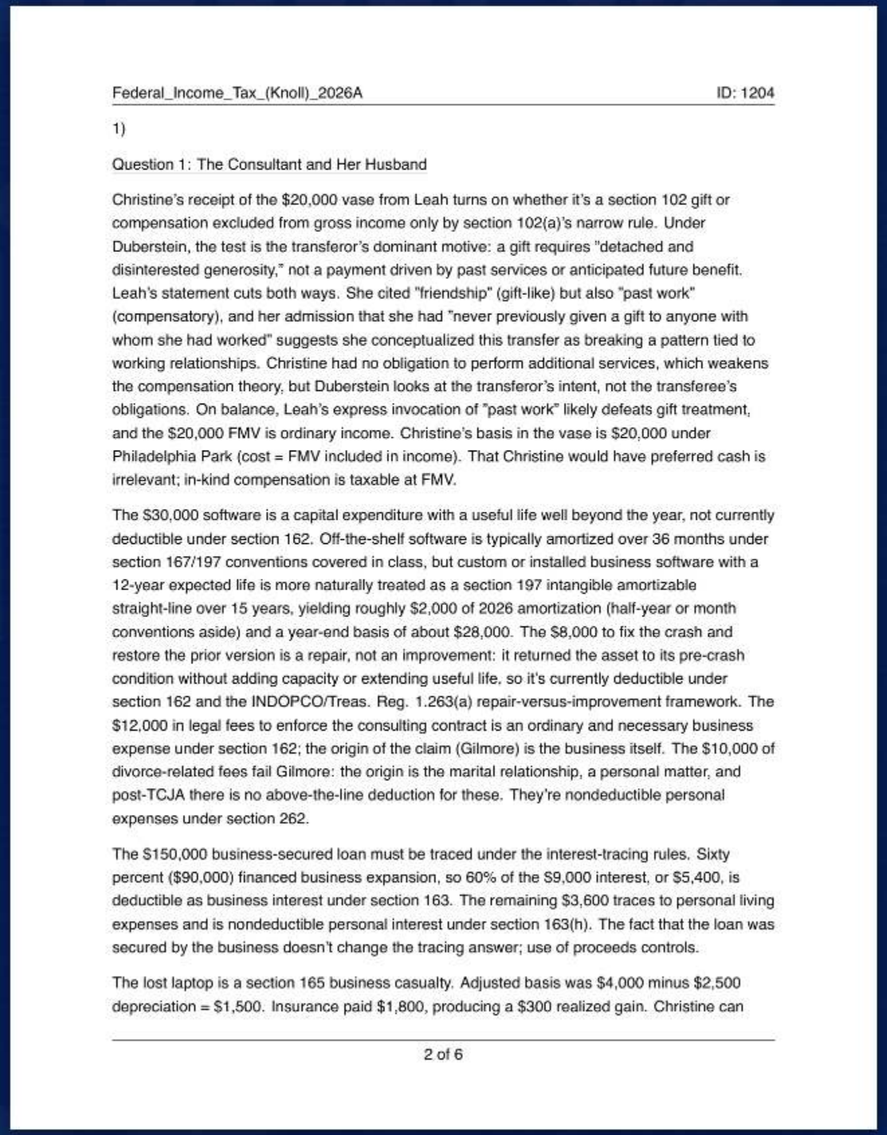
The Prompt

essay_prompt.txt

```
<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
```

- One structured prompt, sent to each isolated cell
- Blocks set the role, the task, and what a strong answer looks like
- The style block bans the machine's tells, including the em-dash
- Citation-honesty forbids invented authority, the single most important rule
- Output is the answer only: no preamble, no headers

Captured, audited, packaged



Federal_Income_Tax_(Knoll)_2026A ID: 1204

1)

Question 1: The Consultant and Her Husband

Christine's receipt of the $20,000 vase from Leah turns on whether it's a section 102 gift or compensation excluded from gross income only by section 102(a)'s narrow rule. Under Duberstein, the test is the transferor's dominant motive: a gift requires "detached and disinterested generosity," not a payment driven by past services or anticipated future benefit. Leah's statement cuts both ways. She cited "friendship" (gift-like) but also "past work" (compensatory), and her admission that she had "never previously given a gift to anyone with whom she had worked" suggests she conceptualized this transfer as breaking a pattern tied to working relationships. Christine had no obligation to perform additional services, which weakens the compensation theory, but Duberstein looks at the transferor's intent, not the transferee's obligations. On balance, Leah's express invocation of "past work" likely defeats gift treatment, and the $20,000 FMV is ordinary income. Christine's basis in the vase is $20,000 under Philadelphia Park (cost = FMV included in income). That Christine would have preferred cash is irrelevant; in-kind compensation is taxable at FMV.

The $30,000 software is a capital expenditure with a useful life well beyond the year, not currently deductible under section 162. Off-the-shelf software is typically amortized over 36 months under section 167/197 conventions covered in class, but custom or installed business software with a 12-year expected life is more naturally treated as a section 197 intangible amortizable straight-line over 15 years, yielding roughly $2,000 of 2026 amortization (half-year or month conventions aside) and a year-end basis of about $28,000. The $8,000 to fix the crash and restore the prior version is a repair, not an improvement: it returned the asset to its pre-crash condition without adding capacity or extending useful life, so it's currently deductible under section 162 and the INDOPCO/Treas. Reg. 1.263(a) repair-versus-improvement framework. The $12,000 in legal fees to enforce the consulting contract is an ordinary and necessary business expense under section 162; the origin of the claim (Gilmore) is the business itself. The $10,000 of divorce-related fees fail Gilmore: the origin is the marital relationship, a personal matter, and post-TCJA there is no above-the-line deduction for these. They're nondeductible personal expenses under section 262.

The $150,000 business-secured loan must be traced under the interest-tracing rules. Sixty percent ($90,000) financed business expansion, so 60% of the $9,000 interest, or $5,400, is deductible as business interest under section 163. The remaining $3,600 traces to personal living expenses and is nondeductible personal interest under section 163(h). The fact that the loan was secured by the business doesn't change the tracing answer; use of proceeds controls.

The lost laptop is a section 165 business casualty. Adjusted basis was $4,000 minus $2,500 depreciation = $1,500. Insurance paid $1,800, producing a $300 realized gain. Christine can

2 of 6

- The intent: make each submission indistinguishable from a student's
- Saved exactly as returned; edits were cosmetic and logged
- Examplify in-class format, with a reserved exam ID
- Not perfect: some differences remained, spacing, font, ID placement
- Those build bugs, not the writing, drove most early detection

Three Findings

Competitive at or Near the Top of the Class

10 / 10
passed every graded exam

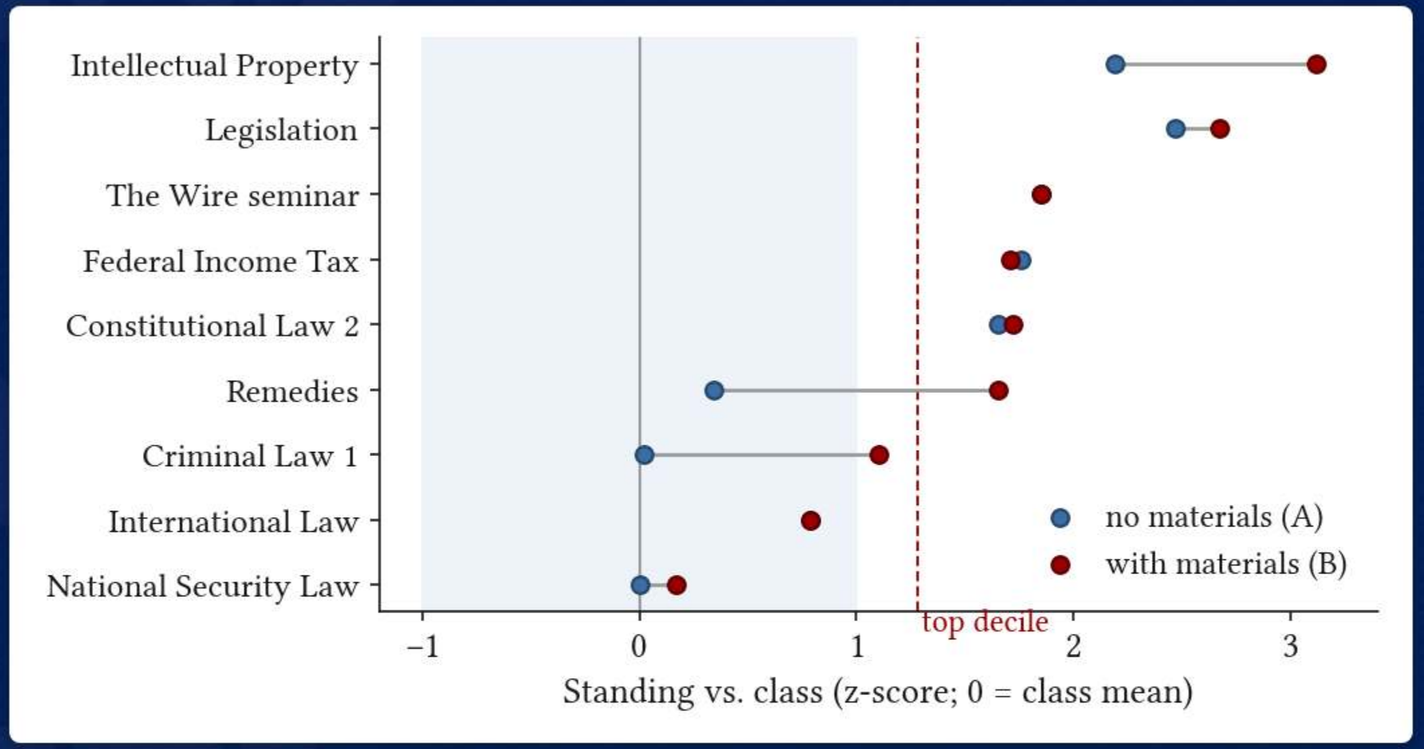
10 / 10
at or above the class median

7 / 10
reached the top decile

2
both arms outscored every enrolled student

One grader declined to grade the AI essays; on one exam, only the MCQs were graded.

Composite Results vs the Students



Each exam, the AI as a z-score vs the class (0 = mean, band = ±1 SD, dashed = top decile). Blue = no materials, red = with materials. Con Law 1 and the Crim Law 2 essay were not graded by the faculty, so are not shown.

AI Results: Highlights

- Legislation: 99pts and 100pts out of 100, above all eighty-four students
- Intellectual Property: the top composite, both arms, in a class of seventy-seven
- The Wire seminar: 40/40, the rubric ceiling the top student did not reach
- Weakest showing in this cohort still respectable, about the sixty-fifth percentile (A- or B+)

The recurring profile: very strong on coverage and issue-spotting, weaker on subtle argument.

Do Materials Help? Multiple Choice

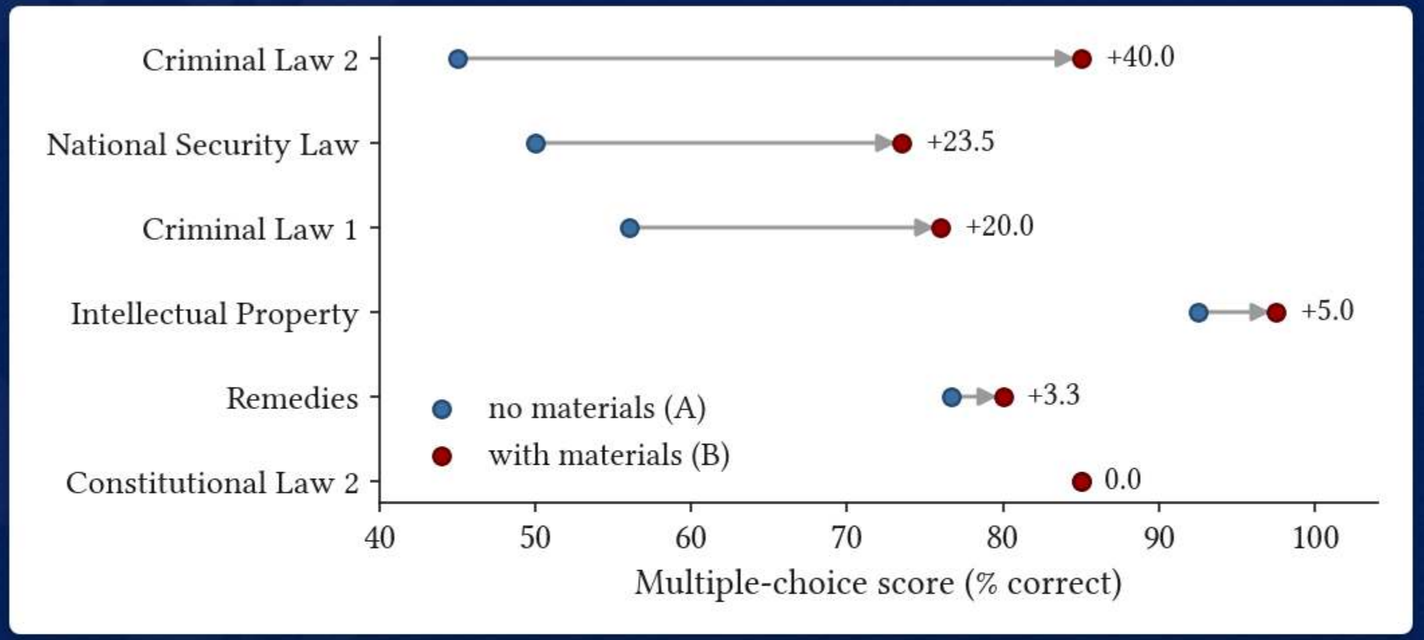
- Large effect
- Average +14 to +15 points
- Up to +40 points (more than 2.5 SD)
- Materials supply the recalled facts

Do Materials Help? Essays

- Almost no effect
- About +0.04 SD across essay-only exams
- Essay competence comes from training, not the outline
- On one exam: MCQ +23, essay -19

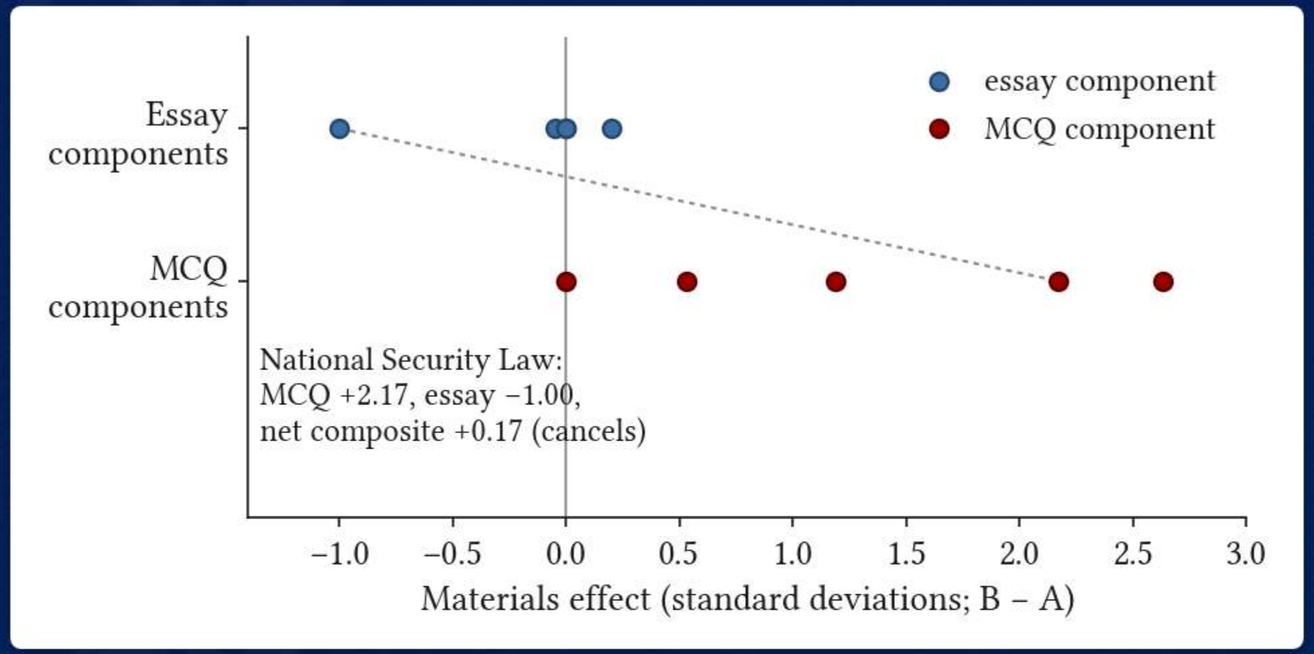
Materials help, but unevenly depending on format and subject area.

Materials and the multiple-choice score



The gain runs from nothing to forty points, largest where the no-materials score started lowest.

The composite hides the action



Materials barely move the essays (near zero) but move multiple choice a lot. National Security Law: the two run opposite and cancel.

Subject Matters as much as Format

Constitutional Law

A body of doctrine the model seems to command fluently from training. This cohort tests it twice, an essay-only section and a multiple-choice-plus-short-answer section, and on both the materials barely moved the score. The no-materials arm reached the top of the class with no course-specific grounding at all.

Can faculty detect the machine?

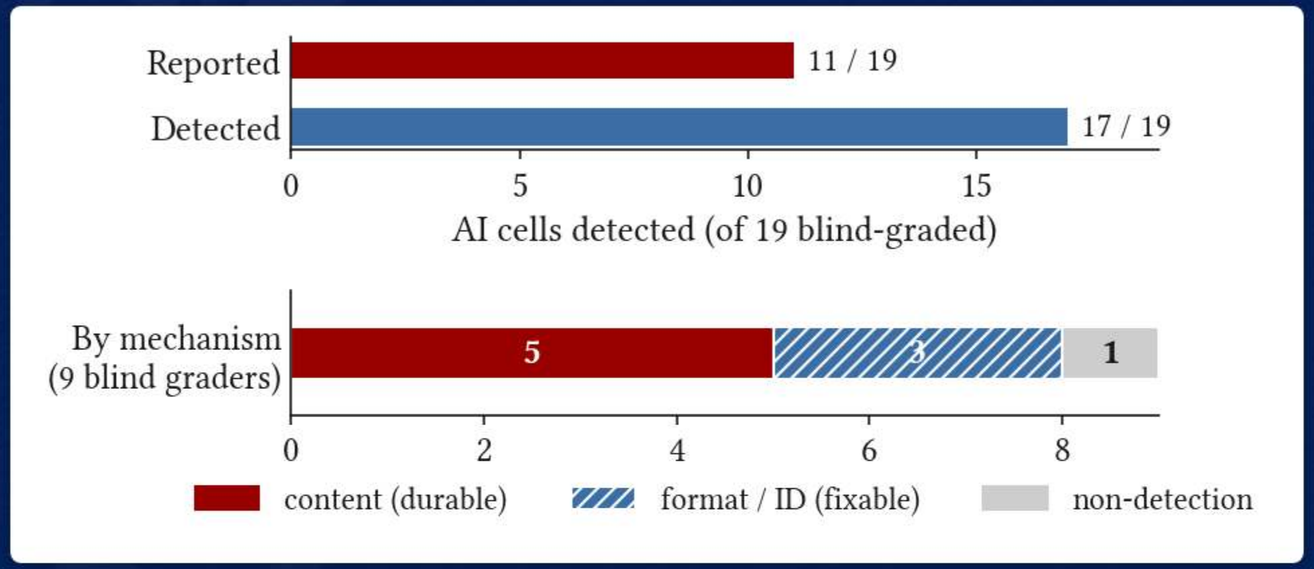
11 / 19
reported (on the record)

17 / 19
detected (caught at all)

5 / 9
graders detected based on content

Most early detection rode on fixable artifacts, not the writing.

Detection: Reported vs Detected



Detected (caught at all) contains reported (put on the record). Mechanism: content (durable) vs format / ID tells (fixable).

Fixable (build bugs)

- Single spacing vs. Examplify's software's double (which apparently varies, too)
- A different body font (which also seems to vary)
- Reserved Exam ID numbers clustered at the end of the set
- A reversed quotation-mark glyph

Durable (the writing)

- Two isolated answers that read as written by one hand
- “The same individual or entity,” three graders said versions of this
- A thoroughness and polish that time-pressured work comprehensively lacks
- One grader declined to grade at all, given confidence in detection

The next wave will fix most of the build bugs; the content signal seems likely to remain.

The Ceiling AI Reaches

Sometimes, the strongest answer was a human's

On the exam where the machine's standing was highest overall, the answer the grader singled out was a student's: a mock judicial opinion, witty and structurally daring. The machine reaches the ceiling of the competent, comprehensive answer; the heights above it, voice, invention, judgment, it does not (yet) reach.

What Are We Measuring?

- When a model that never took your course tops the class, it's no longer clear the exam measures your course
- The exam was always a contrivance built for grading, a weak proxy for real lawyering, and AI only widens that gap
- The urgent ask is intentionality: be clear about what each assessment is meant to measure
- Responses will range, from wholesale redesign to simply naming the mismatch, and reasonable faculty will differ

A problem to sit with, not a verdict.

The next wave, Fall 2026

- Close the blinding leaks: i.e., match font, spacing, and ID placement
- Add models and repeated runs
- Standardize the materials condition at a higher level
- Make faculty detection a measured task

Behind the Paper

Behind the Paper: AI-Assisted Research

Companion to "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

The study and its apparatus

The paper this essay accompanies asks a narrow empirical question: can an artificial intelligence pass your final exam, and does handing it the course materials make it better? To answer that, I had to put a model through a controlled experiment — generate the answers, package them as ordinary Examplify exports, and send them out for blind grading alongside real students. There is an irony here that I did not engineer but cannot ignore. A study of AI sitting law exams was itself conducted with AI as the laboratory. The model was both the subject and, in a different role, the instrument.

This arrangement turned out to be perhaps the most instructive part of the project, and the part the paper itself never discusses. Running a research program with an agent as a collaborator — Claude Code, working from a project folder I controlled — forced me to be precise about something I would otherwise have left vague: what, exactly, can these tools be trusted to do, and where does the trust have to be replaced with structure?

The honest answer is that the machine's role, and its limits, looked different at three distinct stages. The work moved through running the experiment, keeping the record, and writing and verifying the paper. At each stage the agent did something real, and at each stage I had to build a 'fence' around it. This is the story of those fences — and a usable account, I hope, for any colleague who wants to try something similar.

†Michael A. Fitts Professor of Law at the University of Pennsylvania Carey Law School. This essay is a companion to "Can AI Ace Your Exam? A Penn Law Experiment," describing how that study was conducted with the assistance of an AI coding agent. Early working draft (June 2026); faculty and students are anonymized.

1

- A companion essay: this project, done with an AI agent
- Claude Code as collaborator, from a prompt folder I controlled
- Where the tools can be trusted, and where structure must replace trust
- Maybe the most instructive part of the whole project

The question is no longer whether the machine can do what we test,

but whether the exam still measures what we have long assumed it measures.

R. Polk Wagner

pwagner@law.upenn.edu

Project Page

ai-teaching-lab.org/projects/exam-taker/