Penn Carey Law
University of Pennsylvania

Grading Machines — A Summary

A summary of Cope, Frankenreiter, Hirst, Posner, Schwarcz & Thorley, Grading Machines: Can AI Exam-Grading Replace Law Professors?, *Journal of Law and Empirical Analysis (2026)*, DOI 10.1177/2755323X261434265. *Draft posted on SSRN in December 2025:* ssrn.com/abstract=5851362.

What the paper does

Six law professors — from Virginia, Washington University in St. Louis, Boston University, Chicago, Minnesota, and BYU — collected real student exams from four different doctrinal subjects administered at top-30 U.S. law schools. They gave those exams to current-generation language models and asked the models to grade them. Then they compared the AI-generated scores to the human faculty scores.

The design is clean: same students, same exams, same rubrics, different graders. One human; one AI. How close do the grades come?

What they found

With no guidance beyond the grading curve and a basic prompt, the LLM-produced grades correlated with the human-assigned grades at Pearson correlations of up to 0.80 (ranging from 0.66 to 0.80 across the four exams analyzed). When provided with a detailed rubric, that figure reached 0.93 (ranging from 0.78 to 0.93).

For reference: inter-rater reliability among *human* law faculty grading the same exam is itself well below 1.0. Two professors grading the same exam don't always give each answer the same score. An AI-to-human correlation of 0.93, on a problem where the human ceiling isn't 1.0, is striking.

The patterns of disagreement are as interesting as the agreement rate. The AI tended to disagree with faculty in specific directions — systematically under- or over-weighting certain kinds of student reasoning, depending on the subject matter and the rubric's specificity. Those disagreements are where the actual learning sits: they reveal both where AI grading is trustworthy and where human judgment is still doing work the AI doesn't reproduce.

Why this is load-bearing for the course

The study matters less for what it shows about exam grading specifically and more for what it shows about the general claim that AI can't yet reliably do legal-academic work.

If AI-assigned grades correlate with human grades at 0.93 on law school exams — a task that requires sustained attention, domain knowledge, normative judgment, and sensitivity to argumentative structure — then the set of "legal-academic tasks AI can do well" is larger than most faculty assume. Far from perfect correlation, but far from zero. The gap between 0.93 and 1.0 is where the argument is.

The authors are careful about the implication. They are not arguing that AI should replace faculty graders. They are arguing that AI should be used as a supplement: to validate professor grading, to provide substantive feedback on ungraded midterms, to give students feedback on self-administered practice exams. Those are plausible and valuable uses. The deeper point — that AI grading is in the same ballpark as human grading — is the one that destabilizes assumptions about what exams are doing in the first place.

That's why Module IV opens with the exam question.

What to keep in mind

The paper tests current LLMs, not legal-specific AI. A Westlaw Precision or Harvey product specifically trained on legal reasoning might perform better; a general-purpose model without a rubric performs noticeably worse. The 93% number is a high-water mark, not a baseline.

Also: "correlation with faculty grades" is not the same as "correct grading." If the faculty grading is itself inconsistent — and it is — then the AI is correlating with a noisy signal. The paper acknowledges this directly. What it shows is that AI grading is *in the same ballpark* as human grading, not that either is "right."

Finally: the authors are not advocating for full automation. They're arguing for AI as a grading assistant. The line between "AI assists grading" and "AI replaces grading" is where the real pedagogical conversation happens.

One line worth remembering

With a detailed rubric, AI grades correlate with faculty grades at up to 0.93 — on a task where human-to-human reliability isn't 1.0 either. The remaining gap is where the argument about what exams are actually testing gets interesting.

Full citation: Kevin L. Cope, Jens Frankenreiter, Scott Hirst, Eric A. Posner, Daniel Schwarcz & Dane Thorley, Grading Machines: Can AI Exam-Grading Replace Law Professors?, *J. L. & Empirical Analysis (2026), DOI 10.1177/2755323X261434265. SSRN preprint posted Dec. 2025: ssrn.com/abstract=5851362.*