Grading Machines: Can AI Exam-Grading Replace Law Professors?[*]

Kevin L. Cope
University of Virginia

Jens Frankenreiter
Washington University in St. Louis

Scott Hirst
Boston University

Eric A. Posner
University of Chicago

Daniel Schwarcz
University of Minnesota

Dane Thorley
Brigham Young University

Abstract

In the past few years, large language models (LLMs) have achieved significant technical advances, such that legal-advocacy organizations are increasingly adopting them as complements to—or substitutes for—lawyers and other human experts. Several studies have examined LLMs' performance in taking law school exams, finding mixed results. Yet there have been no published studies systematically analyzing LLMs' competence at one of law professors' chief responsibilities: *grading* law school exams. This paper presents results of an analysis of how LLMs perform in evaluating student responses to legal analysis questions of the kind typically administered in law school exams. The underlying data come from exams in four subjects administered at top-30 U.S. law schools. Unlike some projects in computer or data science, our goal is not to design a new LLM that minimizes error or that maximizes agreement with human graders. Rather, we seek to determine whether existing models—which can be straightforwardly applied by most professors and students—are already suitable for the task of law exam evaluation. We find that, when provided with a detailed rubric, the LLM grades correlate with the human grader at Pearson correlation coefficients of up to 0.93. Our findings suggest that, even if they do not fully replace humans in the near future, LLMs could soon be put to valuable tasks by law school professors, such as reviewing and validating professor grading, providing substantive feedback on ungraded midterms, and providing students feedback on self-administered practice exams.

[*] We thank Peter Joy, Jonathan Masur, as well as participants at the 2025 American Law and Economics Association (ALEA) Annual Meeting, the 2025 Midwest Political Science Association (MPSA) Annual Conference, and the 2025 Conference on Empirical Legal Studies (CELS), for helpful comments and suggestions. The Institutional Review Boards (IRBs) of the four academic institutions (University of Chicago, Brigham Young University, University of Minnesota, and Boston University) whose exams and student answers this Article analyzes each determined that the study does not constitute "human-subjects research" as defined by 45 C.F.R. § 46 and therefore does not require IRB review.

Keywords: artificial intelligence, grading, large language models, GPT, legal pedagogy

1 Introduction

Led by Open AI's Generative Pre-trained Transformer (GPT) model family, large language models (LLMs) have achieved significant technical progress over the past several years. These advances have inspired legal technology firms to explore how LLMs might assist lawyers in tasks as varied as document review, contract and motion drafting, and brief writing and editing. In many cases, LLMs are being used to complement the work of lawyers, and—at least for some discrete tasks—they are partly or fully substituting for them. But given the speed of the technology's evolution, it is still unclear if, when, and to what extent LLMs will become a genuinely suitable substitute for human legal analysis.

The legal analysis questions typically administered in law school exams (often referred to as "issue spotters") offer a valuable way to assess this technology's capacity for legal analysis. These exams test foundational doctrinal knowledge, demand creative and multifaceted problem solving, and, importantly, come with a built-in comparison group of human test-takers and graders. For these reasons, several studies have examined LLMs' performance in *taking* law school exams, finding mixed (but increasingly impressive) results (e.g., Fan et al., 2025). Yet there have been no published studies systematically analyzing LLMs' competence in the closely related but conceptually distinct task of *grading* law school exam answers, which lies at the core of legal educators' responsibilities.

This paper analyzes how commercial LLMs, primarily OpenAI's GPT-5, perform in evaluating law school issue-spotter exams. The data comprise four final exams recently administered at four top-30 U.S. law schools—covering three standard first-year courses (civil procedure, contracts, and torts) and one upper-level course (corporations)—and include both the actual student answers submitted and the grades assigned by faculty. We find that, without any guiding instructions beyond a basic prompt and the min–max range for scores, the LLM-produced grades correlated with the actual human-assigned (professor) grades at Pearson correlation coefficients of up to 0.80 (a range of .66 to .80 across the four exams analyzed). When provided with a detailed rubric that the professor used to grade the exam, that figure reached 0.93 (a range of .78 to .93).

We acknowledge that professional regulations, ethical concerns, and general path dependence may prevent law schools from completely replacing human graders with

LLMs in the short term. Indeed, the ethical issues surrounding machine grading echo longstanding debates about automation in other fields, such as autonomous vehicles. A primary concern is that machine graders will make errors that humans would not, raising fairness concerns—even if the aggregate number and magnitude of errors made by LLMs are comparable or lower than those made by humans.

Many of these concerns give insufficient weight to the substantial and well-documented limitations of human grading. Human graders are prone to inconsistency, especially under fatigue. They may introduce unconscious biases based on perceived student identity, even under blind grading conditions. In contrast, AI grading has the potential to reduce these sources of unintended variation and offer more consistent, unbiased assessments. We believe that machine grading should be evaluated against a standard that acknowledges the fallibility of human grading, rather than an idealized standard of human perfection.

Even if AI grading does not immediately supplant human grading, our findings suggest that LLMs could be used for other valuable tasks in law and legal education. For example, LLMs could be deployed to review professor grading for errors or bias and to provide students with rapid, reasonably reliable practice-exam feedback tailored to the grading tendencies of their professors. Outside of the law school setting, AI could serve parallel functions by helping senior lawyers evaluate the work of junior lawyers' and by giving junior lawyers constructive feedback on their own performance. Beyond these human-development applications, reliable AI grading of legal work product could facilitate more consistent benchmarks for assessing AI's legal capabilities and perhaps even operate as one component of a fully automated AI lawyering system.

The rest of this Article proceeds as follows. Section 2 gives a brief history of automated essay scoring and reviews how LLMs have begun to change legal practice and pedagogy over the last few years. It then considers ethical and political issues concerning machine evaluation of exams. Section 3 describes the study's design and methods for measuring how LLM exam evaluation compares with that of human graders. Section 4 provides the empirical results. Section 5 discusses the implications of the findings' for the future of exam grading and the assessment of legal work product more generally. Section 6 concludes.

2 Background and Motivation

2.1 Early Automated-Essay-Scoring Efforts

Automated Essay Scoring (AES) originated in the mid-1960s with the work of Ellis Batten Page, who argued that computers could be trained to score essays as reliably as human graders. In 1968, Page demonstrated this claim with Project Essay Grade, one of the first computerized essay-grading systems. Though early versions were not cost-effective, the proliferation of personal computers in the 1990s renewed the practical viability of AES (Page, 2003).

In 2012, the Hewlett Foundation sponsored the Automated Student Assessment Prize (ASAP). Several teams showed that AES systems could match human rater reliability across multiple essay prompts. Although these initial claims of equivalence were controversial due to methodological flaws, the competition nonetheless spurred interest in modern AES technologies. A year later, EdX, a nonprofit founded by Harvard and the Massachusetts Institute of Technology, introduced an AI software package for grading student essays and made it freely available, with the hope that it would free professors up for other tasks without sacrificing grading validity. EdX president Anant Agarwal claimed that the grading quality is "similar to the variation you find from instructor to instructor" (Markoff, 2013). But critics such as MIT's Les Perelman believed that AES had serious limitations, arguing that its supporters had not provided a valid test that compared the package's performance with that of human graders. Perelman and other educators claimed that computers not only cannot "read," but also cannot assess important aspects of writing such as reasoning, evidence, or clarity (Markoff, 2013).

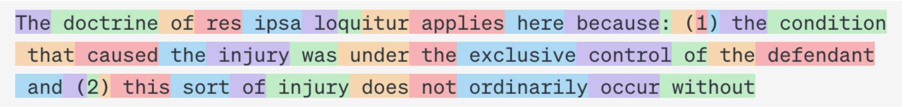
2.2 Large Language Models

The development of a deep-learning architecture called a *transformer* in 2017 began to fundamentally change the capabilities of computers to process text. The transformer was developed by a team of Google data scientists (Vaswani et al., 2017), and became the foundation for a new generation of large language models (LLMs). One prominent family of such models is OpenAI's generative pre-trained transformer (GPT), which applies the transformer architecture together with large-scale pre-training and fine-tuning to natural language tasks.

OpenAI introduced its first GPT model, GPT-1, in 2018. Four years later, OpenAI introduced the general public to LLMs with ChatGPT, which was first based on the company's updated GPT-3.5 model. Many other transformer-based LLMs have since been developed, including Anthropic's Claude, Google's Gemini, Meta AI's LLaMA, and DeepSeek's R1.

At its core, an LLM is a model designed for next-word (or, more precisely, next-token) prediction. During *pre-training*, the model is exposed to vast collections of text, often drawn from publicly available web data or specially curated sources. It learns by predicting subsequent tokens and adjusting its parameters based on the accuracy of those predictions.[1]

After an LLM model is trained, it generates text using a two-step process. First, the model receives text in the form of a prompt from a human user and converts it to a series of *tokens*, that is, a chunk of text such as a letter, subword unit, or word, depending on frequency in the training corpus. This process is known as *tokenization*. For instance, the following statement on the torts principle of *res ipsa loquitur* might be tokenized as the colors indicate:



```
The doctrine of res ipsa loquitur applies here because: (1) the condition
 that caused the injury was under the exclusive control of the defendant
 and (2) this sort of injury does not ordinarily occur without
```

Fig. 1 ChatGPT Tokenization of Statement of Law

Unlike some earlier-developed text analysis methods like the *Bag of Words* (BoW) technique (see, e.g., Juluru et al., 2021)—and critical to generating natural language—the tokens' sequence is retained.

Second, based on the patterns in the training set and the token sequence in the prompt, the model estimates a probability distribution of possible outputs. It produces a set of the next most likely tokens in the sequence, together with an estimated probability for each possible token.[2] In the above example, the possible outputs (the next token in the phrase) might be "recklessness," "negligence," or "intent," for instance. The most likely (and doctrinally correct) output based on well-established *res ipsa loquitur* principles would be "negligence." This second step is repeated as many times as necessary, with the user-provided text plus the previously created output forming the input for the next iteration.

Transformer models such as GPTs, unlike previous LLMs, are innovative in part because they are designed to give greater *attention* to some input tokens than others, based on the strength of relationships with other tokens, as learned during training. In addition, they can process vastly more input tokens than previous technologies (Stollnitz, 2023).

[1] Modern publicly deployed LLMs are also *post-trained*, in which engineers or researchers use various tuning methods to reduce inaccurate, incoherent, or otherwise undesirable language in model outputs. This process can render the models more useful, though it increases discrepancies between model output and how the relevant language in the input corpus is actually used (Choi, 2025).

[2] More formally, the LLM estimates the conditional probability distribution, $P(x_t \mid x_1, x_2, \ldots, x_{t-1})$, for the next token, x_t, given a sequence of tokens, $x_1, x_2, \ldots, x_{t-1}$ (Choi, 2025).

This next-token prediction capability has given rise to models that can process text and generate new content with a fluency that often resembles human communication. Modern LLMs appear to have learned not only statistical patterns of language, but also factual information and conceptual relationships embedded within their training data. As a result, they can perform tasks—ranging from answering complex questions to summarizing dense materials—that were once thought to require human judgment. While their apparent *understanding* remains a product of statistical pattern recognition, the scope and sophistication of these AI models blur the line between automated processing and tasks traditionally entrusted to human experts. This raises a natural question: if these systems can generate useful analyses of texts, might they also be capable of evaluating text quality against pre-specified criteria, such as the standards governing legal work product?

2.3 AI in Legal Analysis, Pedagogy, and Practice

To our knowledge, LLMs have thus far not been systematically tested on law-exam grading, but they have been applied to many other aspects of legal practice and pedagogy.[3] Over just the past few years, researchers have begun to explore how generative language models might aid or replace lawyers in an increasingly diverse set of tasks such as statutory analysis (Engel and McAdams, 2024; Blair-Stanek et al., 2023), cryptosecurities legal analysis (Trozze et al., 2024), contract review (Martin et al., 2024, finding that several LLMs outperformed lawyers in contract-review accuracy), tax problems (Nay et al., 2024), legal-text annotation (Gray et al., 2023; Savelka and Ashley, 2023), the extraction of legally relevant features from corporate governance documents (Frankenreiter and Talley, 2024; Frankenreiter, 2025), bail decisions (Kleinberg et al., 2018), and memo drafting and editing (Simon, 2023, finding that LLMs made significant errors in drafting and editing legal memos).[4] During this same time period, researchers have developed several legal-benchmarking tools to systematically evaluate progress in the legal reasoning capabilities of LLMs (Guha et al., 2023; Fan et al., 2025; Posner and Saran, 2025).

[3] Since transformer-based LLM's introduction in 2018, researchers have also identified myriad ways in which they might complement or substitute humans in non-legal tasks, such as data analysis, creative content generation, code development, information summary and synopsis, decision-making, customer service, and education (see, e.g., Liu et al., 2023; Hadi et al., 2023).

[4] See also Homoki and Ződi (2024); Tiwari et al. (2024); Gamm (2023); Büttner and Habernal (2024); Choi et al. (2025); Schwarcz et al. (2025).

In light of LLMs' performance on legal tasks to date—their successes as much as their failures—commentators have emphasized the need for attorneys to employ these tools ethically and responsibly (Murray, 2023; Pierce and Goutos, 2023). Towards this end, some have called for new regulations governing lawyers (Cyran, 2024), while others have suggested incorporating new rules on the ethical use of LLMs into the rules of professional responsibility (Avery et al., 2023).

LLMs are also changing legal education (Bliss, 2024). Some studies have measured the ability of generative language models to generate answers to law school exams (Choi et al., 2022; Hargreaves, 2023; Blair-Stanek et al., 2024, 2023). One study found that, compared with students using traditional resources, students with access to GPT-4 showed significantly stronger performance on multiple-choice questions but no better performance on complex essay questions (Choi and Schwarcz, 2023).[5] Others have tested LLMs' performance on the U.S. Uniform Bar Examination (Katz et al., 2024; Martínez, 2024), finding that GPT-4 scored above every jurisdiction's passing threshold.

2.4 Exam Evaluation and Grading

Finally, researchers have recently begun to explore LLMs' ability to evaluate and grade test answers across a range of subject matters—including the physical sciences (Kortemeyer, 2023), mathematics (Yang and Zhu, 2022), and collaborative JAVA programming (Tomić et al., 2022)—as well as across different response formats, such as freeform responses (Mitros et al., 2013). Yet no study to our knowledge has attempted to gauge how well generative language models can grade law school exams, which generally require students to identify and analyze legal issues presented in lengthy hypothetical fact patterns (often referred to as "issue spotters").

Evaluating such legal analysis questions poses challenges that most other test formats do not. First, the test questions are in a form known in the testing literature as constructed response, as opposed to so-called *selected-response* answers, such as multiple choice and matching. Constructed response answers are significantly more difficult to validate (Hussein et al., 2019). In particular, most legal analysis questions are extended-response items, the most complex and demanding type of constructed response, for both the test-taker and the grader (Nitko, 1996).[6] Because the answers

[5] Others have examined the psychological effects on students of AI-graded English speaking tests (Chai et al., 2024) and pedagogical effects of AI-graded coding problems (Li et al., 2023).

[6] Many examples of legal analysis questions are available from the Harvard Law School Library's collection of historical examinations from 1871 to 1995, at https://hollisarchives.lib.harvard.edu/repositories/5/resources/4539.

are provided in prose rather than discrete options or code, a sophisticated text analysis is required to evaluate them.[7] Second, and relatedly, there are usually many ways to capably answer legal analysis questions, some of which the grader may not even anticipate. Therefore, algorithmic answer keys are unlikely to produce scores of acceptable accuracy.

For these reasons, AES has historically performed poorly in evaluating the sort of complex reasoning contained in law school exam answers. However, with recent advances in LLM technology, these systems now represent a potentially promising tool to assist or replace human graders in evaluating such exams.

With this background in mind, this Article compares exam-grading outcomes by a state-of-the-art, generally accessible LLM, with those of the professors who authored and officially graded those same exams. For reasons we elaborate below (see Sections 4.3 and 5.1), we believe the primary objective of a machine-grading model should not be to replicate human performance. However, showing that an LLM can approximate the grading outcomes of human experts would at least partly address many of the ethical and institutional concerns about the practice, possibly paving the way for their adoption in law schools and other legal-training environments.

3 Research Design: LLM Performance in Grading Exams

The rest of this Article assesses the ability of LLMs to grade answers to legal analysis questions. In this section, we begin this analysis by describing our research goals and research design choices, identifying the dataset of law school exams we used, and describing the specific LLM grading methods we employed.

3.1 Research Goals and Design

Our overarching goal in designing this study was to determine whether LLMs can be used to make automatic legal analysis scoring both effective and accessible to instructors. Thus, our goal was not to identify the single best approach for grading

[7] For example, in one of the civil procedure questions analyzed here, the question presents a 1,500-word vignette describing a civil cause of action. The test-taker is required to propose responses to the claims on behalf of the defendant. If dismissal is implausible, the test-taker must analyze how the case could instead be heard in federal court. To receive full credit, the test-taker's answer must address timing of amendments, improper joinder, dismissal via Federal Rule of Civil Procedure 12(b) (i.e., personal jurisdiction, venue, service/process, failure to state a claim), removal to state court, and general strategy (e.g., whether dismissal and/or joinder is even preferable). For each issue, a test-taker would receive partial points for identifying a given procedural move and full points for correctly applying the appropriate rule and reaching the most-likely conclusion.

exams with LLMs, but rather to evaluate the ability of current LLMs to validly grade exams across a variety of methods that instructors could easily implement using information they likely already have at hand.

To accomplish this goal, we task LLMs with grading student exam answers from four law school courses, and compare the resulting scores to the grades assigned by the respective instructors of those courses. One way of thinking about this approach is to treat an exam grade as a noisy signal of the student's latent (unobserved) mastery of the subject matter (Chilton et al., 2024). In the terminology of the applied econometrics literature on prediction and machine learning (Kleinberg et al., 2015; Mullainathan and Spiess, 2017; Athey and Imbens, 2019), this can be viewed as a prediction or signal-extraction problem: the LLM seeks to recover a latent trait (a student's true mastery) from the observable text of their answer. Employers and others rely on this signal when evaluating law students. Thus, for instructors designing assessment, the primary goal is to maximize the quality of this student-ability signal. As the quality of the signal deteriorates, GPA becomes less informative, leading employers who rely on GPA to make worse hiring choices.

A challenge for evaluating the effectiveness of automated legal analysis scoring is establishing a baseline for the evaluation. Because humans are also imperfect graders (as we discuss further in Section 5.1 below), we cannot conclude that the human evaluators' decisions perfectly reflect a student's true abilities. Thus, discrepancies between LLM-generated and human-assigned grades do not imply that the *machine*'s grades are invalid; mistakes by human graders could also be the source of any discrepancies. Conversely, observed agreement between the two types of graders does not imply validity, as both could be affected by measurement error. More generally, patterns of agreement or disagreement alone do not allow us to determine which of any two given evaluators—such as human and machine—has produced the more valid measure of that student ability.

For the purpose of our project, we resolve this problem pragmatically, by assuming that the goal of automatic legal analysis scoring is to replicate the human grader's evaluations as closely as possible. This assumption is an appropriate goal for many of the tasks automatic legal analysis scoring could be used for.

We believe this assumption is defensible for at least three reasons. First, no 'true' measure of student ability is available. Second, human grading is currently the only accepted practice. Therefore, in the minds of policy makers (such as law school deans, or the American Bar Association) and other stakeholders, a model that approximates human-grading outcomes will surely be considered more legitimate, all things equal, than one that deviates from those outcomes. Third, for applications such as providing feedback to students on practice exams, users might be more interested in scores approximating the human grader's grading than in 'true' measures of student ability. We return to this issue in Section 4.3, where we empirically compare agreement between LLM- and human-generated grades, on one hand, to agreement between human-generated grades produced for the same exam at different periods, on the other.

3.2 Dataset

We began by assembling a dataset comprising final-exam questions and anonymized student answers from several core law school courses, including three traditional first-year courses (civil procedure, contracts, and torts) and an upper-level course (corporate law). Each exam was recently authored, administered, and graded by one of the authors, each at a different U.S. law school ranked among the country's top 25 (according to the *U.S. News & World Reports 2025 Law School Rankings*). The four exams differ by subject matter and length, but each follows the legal-analysis-question format familiar to U.S. law students and instructors (see, e.g., footnote 7), and each exam included at least two legal analysis questions. Each instructor relied on a grading rubric in evaluating their exams, the complexity of which varied among the instructors.[8]

3.3 LLM Grading Method

All reported evaluation scores were obtained using OpenAI's GPT-5 model.[9] Queries were submitted through the OpenAI Batch API via the *openai* package in Python. This setup allowed us to process thousands of prompts efficiently and at scale while maintaining a consistent prompt structure.[10]

[8] Some of the exams also contained multiple-choice questions or policy essays not analyzed here.

[9] In unreported analyses, we compared the performance of the GPT-5 model with OpenAI's earlier models as well as leading models from Anthropic's Claude and Google's Gemini families. We find broadly similar results across model families, with reasoning models generally outperforming flagship models and newer models generally outperforming older ones. Because our aim is not to conduct a model horse race or identify the single best approach to exam grading, we leave a systematic comparison of model families to future work. For present purposes, it is sufficient to note that we did not identify any alternative model capable of producing substantially stronger results than those reported here without compromising our goal of keeping the prompting strategies straightforward and accessible to instructors.

[10] We acknowledge that not all potential adopters of our approach will have the technical means to use the API. Instead, we anticipate that many instructors will rely on ChatGPT's chat interface. However, there is no *a priori* reason to expect substantially different results when using the chat interface, provided users ensure that their prompts are uniformly structured. To achieve this, it

We used various prompting approaches to prompt LLMs to evaluate the exam answers. They ranged from simple prompts asking the model to assign points based on its understanding of the law, to more complex prompts that incorporated instructor-supplied grading rubrics with detailed scoring instructions. We also tested approaches in which the LLM conducted pairwise comparisons of exam answers, with human researchers subsequently deriving grades using a scaling model based on those comparisons.

The four approaches to automated legal analysis scoring are listed below. All exams involve multiple independent questions based on factual scenario; we generally obtained scores for the answers to each question and calculated total scores for each student's exam by summing those question scores. Some approaches also involved scoring individual elements of an answer ("issues," in law school parlance) and then summing those element scores to calculate a score for the question.

Open: Under this prompting approach, the model receives a simple prompt directing it to assign a numerical score for the answers to each question based on the text of the question, the student's answer, and the maximum score for the question. This approach relies solely on the internal capacity of the LLM to interpret and apply legal concepts in grading. It has the benefit of simplicity and ease of adoption, though, since it draws its information entirely from external sources, it is most likely to deviate from the professor's view of the ideal response.

Rubric: This prompting approach provided the model with the same information as the *Open* approach and likewise asked the model to produce a score for the answers to each question, but also incorporates into the prompt the rubric information that the professor used in grading the exams. Although the rubrics vary in different ways, they all outline the elements that answers should address for each question, along with maximum points allocated to each element. The prompt asked the model to consider those elements and points for each element in determining the score for each question, though did not require the model to include those in its output. This approach provides the advantage of nudging the model to consider the information that the professor values, but it is more human-labor-intensive than the open approach, as it requires the instructor to create a rubric for each exam question.

Bespoke: This approach follows the *Rubric* approach, but included some exam-specific language and instructions requiring the model to output disaggregated scores for each rubric element. Question scores are obtained by summing the scores

may be helpful to prepare prompts in a separate word or text document before copy-pasting them into the interface.

that the model assigned to the answer for each rubric element. This approach amplifies both the advantages and disadvantages of the *Rubric* approach, as it is more guided but also and more labor-intensive, as parts of the code must be customized for each different rubric.

Pairwise: This approach is similar to the *Rubric* approach in that the model is provided with the exam question, student answers, and the rubric. However, rather than asking the model to return a score for each exam, each query presents two students' answers, and the model is asked to conduct pairwise comparisons between the two and determine which of them is better.[11] After obtaining pairwise comparisons for all answer pairs, we compute scores for each answer using the Bradley–Terry scaling algorithm (e.g., Hunter, 2004), and then rescale those scores to match the range of scores assigned by the instructor. This approach shares most of the advantages and disadvantages of the *Rubric* approach, but it is more complex, substantially more computationally intensive, and therefore much more expensive to implement. On the other hand, as we discuss in more detail in Section 4.2.2 below, there is reason to expect that it might produce scores that more closely track the underlying construct of interest.

* * *

We used each of these four methods to produce scores for each exam answer for each of the four subjects—civil procedure ($N = 53$), corporations ($N = 36$), contracts ($N = 66$), and torts ($N = 50$)—using each of the four methods above. The scores derived from each of the four methods were each compared in turn with the scores that the respective human grader assigned. Our primary measures of comparison are the Pearson correlation coefficient (r) and the Spearman rank correlation coefficients (ρ), both calculated between the human- and machine-graded total scores for each student.

Both measures capture different aspects of agreement and deviation between human- and machine-assigned grades, and which is more appropriate might depend in part on prevailing grading practices. The Pearson correlation coefficient captures the relative distance between raw exam scores, making it well suited in the context of grading practices that place significant emphasis on raw scores for allocating exams to different grades (e.g., letter grades like A, A-, or B+). By contrast, insofar as grading practices rely primarily on ranking exams and assigning grades based solely

[11] In preliminary evaluations, we found that LLMs can exhibit position bias, favoring either the first or second answer in a prompt. We therefore compare each answer pair twice, once in each direction, in order to eliminate the impact of any such bias on the question scores.

on rank (for instance, through curve bins with predetermined numbers of students assigned to each grade category), information about score distances becomes less relevant, and rank order becomes paramount. Spearman's rank correlation is better suited in this context, as it focuses on the relative ordering of students and is unaffected by the distribution of the scores assigned to them.

4 Results

4.1 Main Results

Figure 2 provides a high-level overview of the Pearson and Spearman correlations by exam and prompt approach. Across all exams, our most straightforward approach, *Open*, achieves impressive performance: The Pearson correlation between LLM- and instructor-assigned grades is 0.66 or higher for all exams, with the most highly correlated exam reaching 0.80 ($\mu = 0.71$). The corresponding Spearman correlations are slightly lower, ranging from 0.59 to 0.77 ($\mu = 0.67$). These results suggest that LLMs, through their training alone and without further instruction, are able to capture meaningful differences in exam quality.

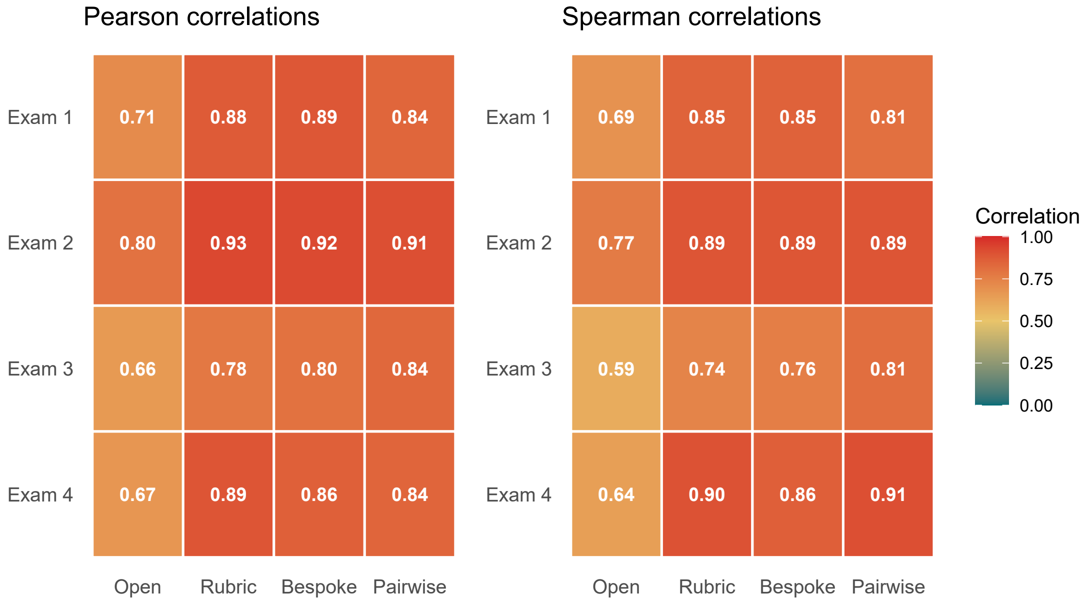


Fig. 2 Heatmaps of LLM performance across exams. The left heatmap shows Pearson correlations between LLM- and instructor-assigned grades for each exam and grading approach, the right one shows the corresponding Spearman correlations.

Across all four exams, grading accuracy improves meaningfully under the *Rubric* approach, in which an instructor-created rubric is added to the information provided to the LLM in the *Open* approach. Pearson correlations for the *Rubric* approach range from 0.78 to 0.93 (μ = 0.87), Spearman correlations from 0.74 to 0.90 (μ = 0.85). Interestingly, there appears to be some variation in how much the switch from *Open* to *Rubric* improves performance. For example, LLM grading of Exam 4 performed roughly similar to Exam 3 under the *Open* approach, but improved considerably more under the *Rubric* approach. This difference may reflect variation in the amount of additional information contained in the rubrics themselves, or conversely, variations in the model's ability to evaluate answers to certain exams without such information.

Finally, the *Bespoke* and *Pairwise* approaches both offer some potential improvement over *Open*, but do not appear to consistently improve performance in comparison to *Rubric*. We revisit both of these results in more detail in Section 4.2 below.

Figures 3 and 4 display the same results in more detail. Figure 3 contains scatterplots showing how LLM-generated grades (on the y-axis) compare to human-assigned ones (on the x-axis), both scaled to 100. The gray dashed lines in each chart indicate a hypothetical perfect agreement: If all points fell on this line, the model would have replicated the human-assigned grades perfectly. Points that appear above the dashed line represent exams that received a higher relative grade from the LLM than from the human grader, whereas the opposite is true for points that fall below the dashed line. The solid colored line in each plot displays a linear regression line fitted to the observations.

Notably, for three of the four exams, the model appears to systematically overshoot or undershoot the human-assigned grades. This pattern emerges consistently under the *Open*, *Rubric*, and *Bespoke* approaches, although it appears to play out differently in *Pairwise*.[12] The main performance metrics we rely on do not account for this over- and undershooting, since most law schools grade on a curve. What matters is relative, not absolute, performance. Accordingly, when interpreting the scatterplots, the more important question is whether the points are tightly clustered around the regression line, not whether they are clustered around the dashed line indicating perfect agreement.

The relative ordering of correlations in Figure 2 is also evident in the scatterplots: the points are more tightly clustered around the regression line under the *Rubric* approach than under the *Open* approach, indicating that the model generates a closer approximation of human grading in a relative sense (even if absolute scores

[12] It deserves mentioning that the Bradley-Terry method used to calculate scores from the pairwise comparisons does not naturally produce scores on the same scale as the original grades. The resulting scores therefore require rescaling, which necessarily involves an arbitrary scaling choice.

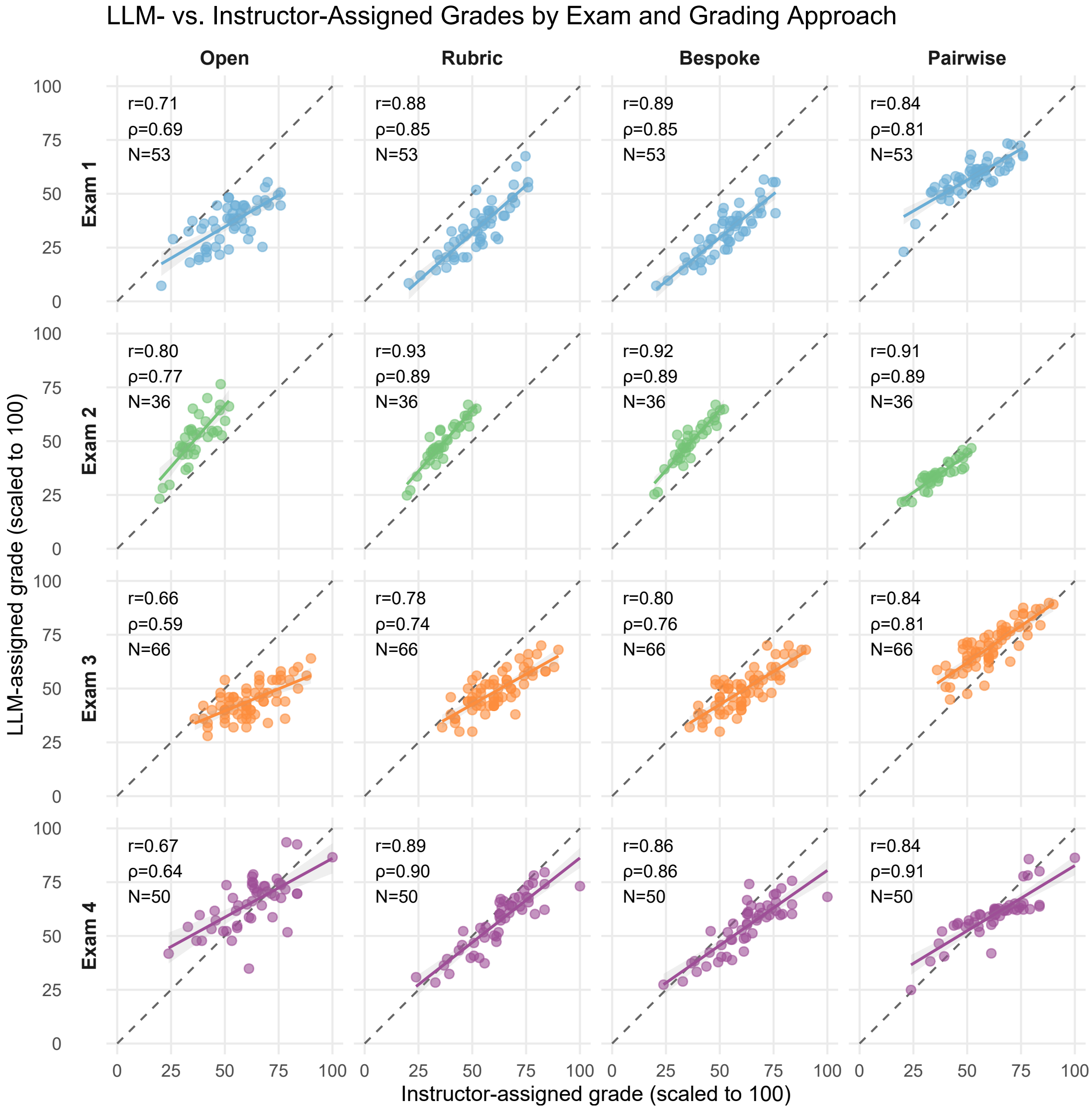


Fig. 3 Comparison of performance across grading approaches. Each scatterplot plots LLM-assigned grades against instructor-assigned grades for one exam, with the dashed line indicating perfect agreement. The solid line denotes a linear regression line fitted to the observations. Numbers in the upper-left corner of each panel report the Pearson correlation (r), Spearman correlation (ρ), and number of observations (N).

are systematically a bit higher or lower). Results from *Bespoke* appear broadly similar to those of *Rubric*, not only in terms of correlations but also in the overall grade distribution. By contrast, while *Pairwise* performs comparably to *Rubric* and *Bespoke* in terms of correlations, its grade distribution differs noticeably.

Figure 4 breaks down the exams into individual question-level scores, displaying the correlations for each question within each exam, along with the correlation for the total score for each exam. The different bar colors represent the four prompting approaches. As with the overall exam scores, the *Rubric* approach substantially outperforms *Open* for almost all exam questions. The only exceptions are a few questions in Exam 3, where *Open* achieves comparably strong results to *Rubric*. These questions appear to be outliers, as the performance of all models on these questions is substantially lower than for other exam questions. Consistent with the results for total scores, *Bespoke* and *Pairwise* rarely yield meaningful improvements over *Rubric* at the individual question level.

Within individual exams, the correlations for total scores are generally similar to or higher than those for the questions on which LLM grading performs best. This pattern suggests that there is noise in either the human or LLM grading, or both, that is canceled out when a larger number of exam questions are aggregated into a total score.

4.2 The More Resource-Intensive Approaches

The results in the previous section show that the *Rubric* approach outperforms *Open* by a substantial margin across both exams and questions. We now turn to consider in more detail whether the LLM's performance can be improved by using more labor-intensive approaches, either *Bespoke* or *Pairwise*.

4.2.1 Bespoke Rubric Output

Although *Rubric* provides the model with explicit guidance on how to value the different components of a response, it ultimately relies on the model to adhere to those instructions by calculating final question scores based on the partial scores it assigns to the sub-topics specified in the rubric. The *Bespoke* approach differs in that it requires the model to output scores at a more granular level, which we then aggregate to generate a question-level score. This approach essentially mirrors that of an instructor who uses a spreadsheet to record the number of points a student earns for each element of their answer and to sum them into a score for the question.

A priori, we expected the *Bespoke* approach to perform better than the *Rubric* approach, as it compels the model to follow the instructor's rubric more closely. By

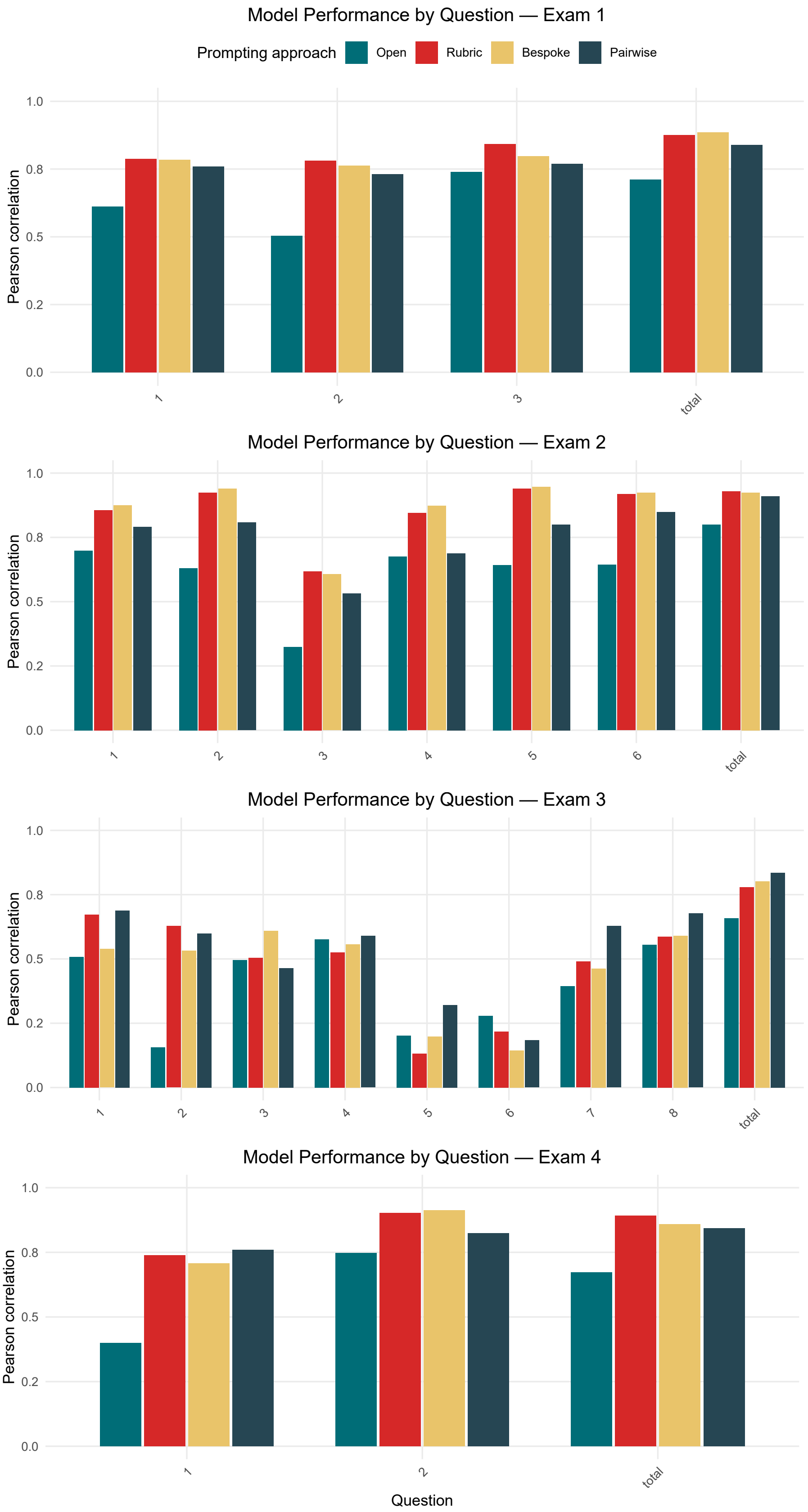


Fig. 4 Correlations between LLM- and instructor-assigned grades at the question and exam levels. Barplots show Pearson correlations for each individual question score and total exam scores for each exam. Colors indicate the prompting approach.

contrast, the *Rubric* approach relies on the model's own reasoning to interpret and apply the rubric's instructions, without any assurance that it will do so faithfully.

The *Bespoke* approach entails significant costs. Specifically, it requires more effort in prompt creation than *Open* or *Rubric*. The same prompt template can be used for all exams under the latter two approaches. However, because the number of topics, and the points assigned to each, typically varies across exam questions, the prompts and code for the *Bespoke* approach must be customized for each exam.

The third column of plots in Figure 3 above shows how grades obtained under *Bespoke* compare to instructor-assigned grades. Pearson correlations coefficients range from 0.80 to 0.92 ($\mu = 0.87$). Where *Bespoke* yields better correlations than *Rubric* (Exams 1 and 3) the improvements are small (0.01 and 0.02), and in the other two exams *Bespoke* performs worse than *Rubric*. Figure 5 illustrates how the grades obtained using the *Bespoke* approach (y-axis) against those obtained using the *Rubric* approach (x-axis). These plots confirm that the grades generated using each of these approaches are closely aligned, with Pearson correlations from 0.94 to 0.99.

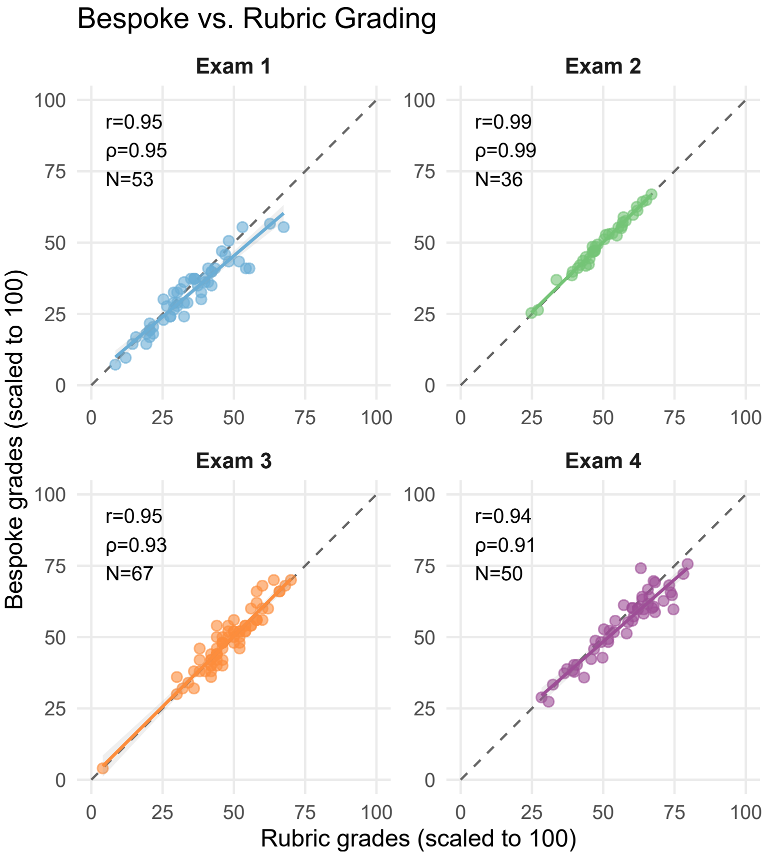


Fig. 5 Comparison of performance between *Bespoke* and *Rubric*.

These results suggest that GPT-5 is able to implicitly implement the element-level scoring contained in a rubric without the additional assistance of a prompt that explicitly forces the model to output scores for each element. A more practical conclusion is that the more labor-intensive prompts and code required for the *Bespoke* approach is likely not necessary in order to obtain reliable grades from a state-of-the-art LLM.

4.2.2 Pairwise Comparisons

Our second attempt to improve LLM grading relies on pairwise comparisons instead of grading on point scales. We use pairwise comparisons because, in other contexts, studies have found that they often produce more reliable ratings than numerical or absolute scales (such as Likert ratings). In areas ranging from coding of latent features of political texts (Carlson and Montgomery, 2017) to computerized tomography (CT) image-quality evaluation (Hoeijmakers et al., 2024), pairwise comparison tends to yield higher reliability and finer-grained distinctions than traditional scales. The likely reason for this advantage is that forcing direct comparative judgments reduces scale-interpretation variance and central-tendency bias. In addition, comparing items promotes effective discrimination, even when absolute standards are ambiguous.

However, pairwise comparison methods are analytically intensive, as the number of comparisons required grows quadratically with the number of total answers to grade. For instance, a class with 67 student exams, like Exam 3, requires 2,211 pairwise comparisons per question.[13] This likely places the approach out of reach for users who intend to implement LLM grading manually via a chat interface, and it renders the approach more expensive by an order of magnitude for those who are able to use it. Another limitation of pairwise comparisons are their inability to deliver absolute scores anchored in a predefined scale (Hoeijmakers et al., 2024), although the output scores can simply be rescaled to the desired scale.

The fourth column of plots in Figure 3 shows how grades obtained under *Pairwise* compare to instructor-assigned grades. Like *Bespoke*, *Pairwise* offers at best marginal improvements over *Rubric*. The Pearson correlations coefficients range from 0.84 to 0.91 ($\mu = 0.86$). However, its grade distribution differs markedly. Figure 6 plots the *Pairwise* total scores for each exam (y-axis) against the total scores obtained using *Rubric* (x-axis).

[13] As discussed above, because of the possibility of position bias, we run each comparison in both directions, resulting in 4,422 comparisons for each question in Exam 3.

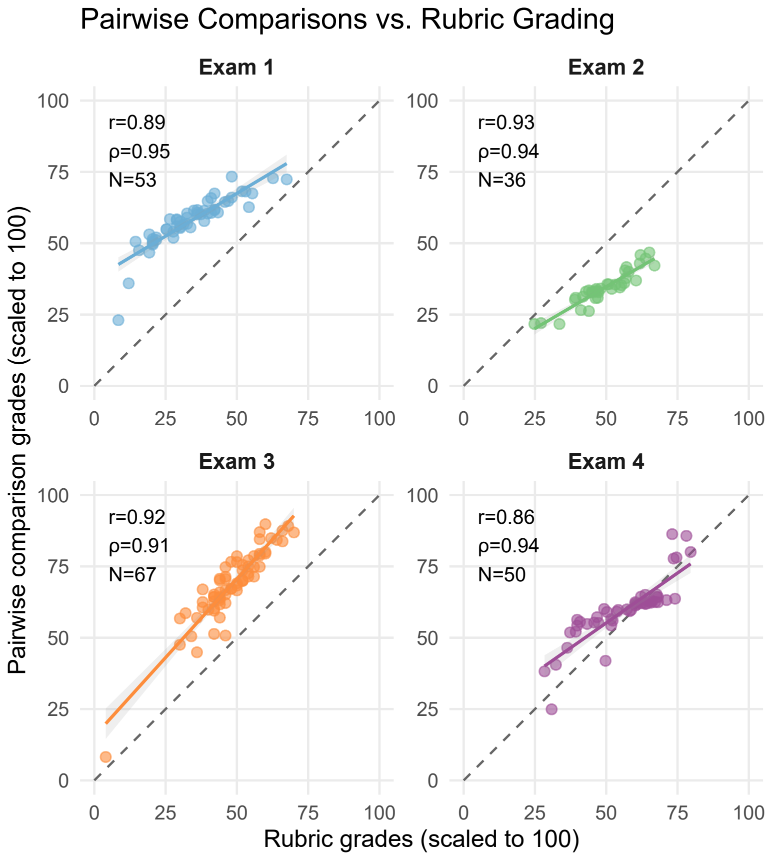


Fig. 6 Comparison of performance between *Pairwise* and *Rubric*.

Figure 6 further illustrates how grades generated under *Pairwise* diverge from those produced under *Rubric*. Although the *Pairwise* grades are less closely aligned with *Rubric* than *Bespoke* grades are, they remain more closely aligned with *Rubric* than with instructor-assigned grades. Taken together, these patterns suggest that all three approaches—*Rubric*, *Bespoke*, and *Pairwise*—are responding to similar features of the student answers, including components that are weighted differently or not fully reflected in human grading.

Overall, these results indicate that our more labor- and resource-intensive attempts to improve performance relative to *Rubric* did not significantly improve the grading performance of the LLM. In other words, the relatively low-effort and easy-to-implement *Rubric* approach already yields performance that appears difficult to improve meaningfully upon using simple prompting methods.

4.3 The Benchmark Question

Our results indicate substantial convergence between the LLM-generated exam scores and those assigned by the original instructor, with our preferred approach yielding Pearson correlation coefficients between 0.78 and 0.93. In other words, the

LLM scores correlate at high-to-very-high levels with human scores, but not perfectly. What implications do these results have for the ability of LLMs to substitute for instructors in grading student exams?

An important consideration in attempting to answer this question is the appropriate benchmark against which to evaluate the accuracy of LLM-generated scores. As alluded to above, human grades are likely themselves noisy signals of the quantity of interest, namely, answer quality. One way to approximate the degree of noise in human-assigned scores is to examine variation across different grading instances in how a human grader scores student answers, i.e., intra-coder variability. If a human grader exhibits meaningful variation across instances, then a natural benchmark for evaluating LLM grading performance is whether LLM-generated scores fall within the range of disagreement implied by the human grader's own intra-coder variability.[14]

Measuring intra-coder variation involves substantial challenges.[15] In most cases, exams are graded only once. Regrading an exam is complicated by two possible measurement concerns. First, the instructor may remember their initial assessment, biasing the second evaluation toward the conclusions reached by the first. In that case, the measured intra-coder variation would be biased toward greater agreement. Second, a regrading exercise inherently occurs under lower stakes than the initial grading, possibly biasing the intra-coder variation toward disagreement.[16]

Despite these challenges, we attempted to obtain insights into benchmark, that is, how the correlation between LLM-assigned and instructor-assigned scores, on one hand, compares to intra-coder variation, on the other. For this, one of us regraded a portion of an exam included in this study (Exam 1, Question 2). Figure 7 presents these grades and compares them both to the original instructor-assigned grades and to the LLM-generated grades obtained under the *Rubric* approach. The two sets of instructor-assigned grades are more closely aligned with each other than either is with the LLM-generated grades, suggesting that, at least in this case, the

[14] This concept differs from the concept of *inter*-coder reliability, which is often measured in research involving human graders. Inter-coder reliability is a less appropriate benchmark here because ABA rules require that law school exams be graded by the instructor who taught the class, and very few exams are shared across instructors.

[15] Chilton et al. (2024) evaluate the reliability of human grades by comparing the consistency of their exam grades across different courses for particular students with the aggregate exam grades that the student receives over their second and third years. A similar exercise could be used to compare LLM and human grading.

[16] In addition, there is the formidable practical issue of convincing human graders to regrade their student examinations.

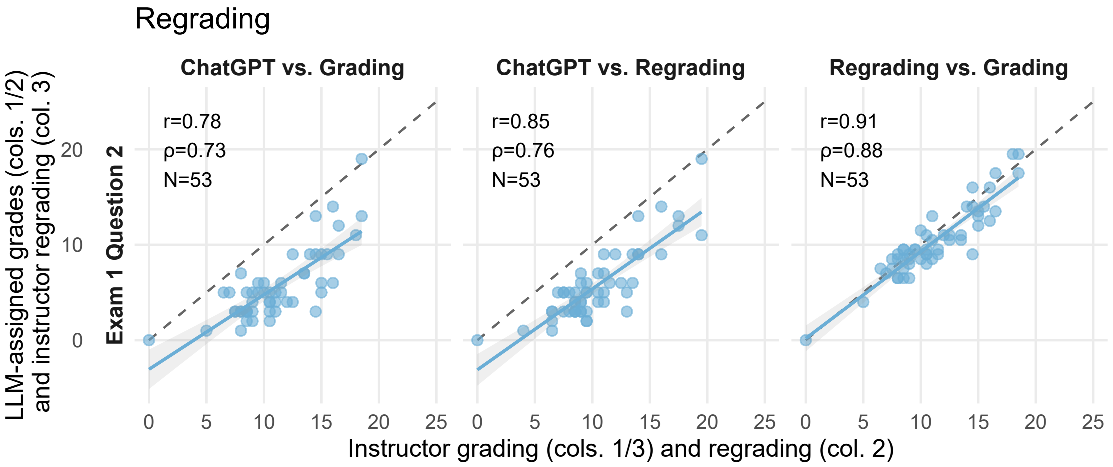


Fig. 7 Comparison of grades obtained from two rounds of grading the second question of Exam 2 by the same instructor.

divergence between LLM and instructor grading exceeds the level of intra-coder variation. We note, however, that the correlation between LLM-generated scores and human-generated scores in some other exams and questions—particularly those in Exam 2—surpasses the observed intra-coder variation for Exam 1, Question 2.[17] This exercise suggests that LLM-human variation and intra-human variation have the same order of magnitude, and that the former is roughly comparable to naturally occurring intra-coder variation.

4.4 Differences in Exam Structure and Rubric Detail

Although our sample size of four exams is necessarily limited, the results suggest that greater rubric specificity meaningfully enhances the model's ability to replicate human scores. The rubrics for these exams vary substantially in their level of detail, providing a useful basis for examining this relationship. For example, Exam 2) employed a highly granular rubric assigning binary (0/1) scores to each element, while Exam 3 relied solely on broad, undifferentiated question-level guidance. Exam 2 produced the strongest correlation between human and LLM-assigned scores (0.93), whereas Exam 3 yielded the weakest (0.78). The rubrics for Exams 1 and 4 fell between these two extremes in terms of specificity, and for these exams the LLM produced correspondingly intermediate correlations (0.89 and 0.88). These

[17] We also note that the regraded scores are more highly correlated with the LLM scores than the original grades were correlated with the LLM scores. While this difference may not be statistically significant, it may reflect the fact that the regrading effort took place after the initial results presented in this paper had been compiled and read by the grader.

patterns suggest that the precision and structure of grading rubrics may be an important determinant of how effectively LLMs are able to emulate human grading. However, given our limited data, this remains a hypothesis and warrants further investigation.

5 Discussion and Implications

This study's results indicate that currently-available LLMs have the capacity to replicate human grading of law school exams with a high degree of accuracy. When the prompt architecture included a highly-detailed rubric that a human grader used in their process, OpenAI's most recent LLM, GPT-5, was able to achieve a Pearson correlation of up to 0.93. This is especially impressive given that even a human who regrades their own exam is unlikely to achieve a perfect correlation, as discussed above.

5.1 Automated Legal Analysis Grading

Despite the impressive accuracy shown by LLMs in grading exams, instructors and legal education programs hoping to substitute human grading with machine grading are likely to face both ethical questions and institutional hurdles. For instance, even if machine-grading correlates highly with human grading, machines may make mistakes that human graders would not, resulting in some students receiving different (though not necessarily 'wrong') grades. Such mistakes may or may not be random; they might, for example, reflect biases in the data on which they were trained. For instance, although we found no evidence of this in our study, it is possible that LLMs could systematically penalize exams that use particular words or sentence structures associated with certain types of students or responses, even when those features are legally irrelevant and might not affect a human grader's judgment (Barocas and Selbst, 2016).

More practically, students and prospective employers may place less trust in machine-generated grades than in human grades, even when the results are close to identical, reflecting the well-documented tendency of humans to undervalue output they believe to be AI-generated rather than produced by a human (Harasta et al., 2024). In addition, some may argue that instructors are hired and paid in part to evaluate students, and thus it is ethically inappropriate for them to delegate that responsibility to machines.[18] Those making such arguments may point to the facts that law students are generally required to complete exams without AI assistance

[18] We note that a similar issue arises with the issue of instructors delegating grading to teaching assistants.

and (at least for now), it is human judges to which they will eventually need to present arguments.

Although many of these concerns about fully automating legal grading are valid, they should be balanced against important countervailing considerations. Most notably, the potential shortcomings of machine graders should be considered alongside the well-documented limitations of human graders. Human grading should not be assumed to represent an infallible gold standard, such that any divergence from human scores necessarily signals error. To the contrary, human grading is itself deeply flawed.

One of the most pervasive shortcomings of human grading is inconsistency: evaluators often assign different scores to similar answers, even when they use detailed rubrics (Liew and Tan, 2024; Yang et al., 2020). Scoring inconsistencies may be especially prevalent when grading is spread over long periods or conducted under varying conditions that lead to grader fatigue (Kumar and Boulanger, 2020). These conditions are hardly uncommon in the law school context, as law school essay grading is both labor-intensive and time-consuming. Grading fatigue can likely be mitigated through breaks, but not eliminated altogether. Law school instructors often grade a substantial number of exams, each consisting of many pages, a task that can take multiple days. In fact, this problem was evident within the data we used for this experiment. Upon a manual review of the individual exams with the largest disparities between the assigned and AI grades, we identified some instances in which a human grader appeared to have made straightforward grading errors. In several cases, the human grader appears to have awarded more points than was justified under their own rubric. Given that error and inconsistency are inherent shortcomings of all human grading, such grading is a highly imperfect proxy for evaluating the true capacity of LLM grading.

Moreover, even when grading under ideal conditions, well-intentioned and expert human graders bring prior experiences, expectations, and unconscious biases to their evaluations of student exams. Essay exam scores thus reflect not only the strength of the substantive arguments in the exam answer, but also the identity of the graders and their beliefs about the student who produced them. For example, experimental evidence shows that evaluators are prone to “halo bias,” where positive impressions in one domain—such as classroom performance or even grammar and style within the exam answer—result in higher scores, although they are ostensibly irrelevant under the formal grading criteria (Malouff et al., 2014). Other studies of writing assessments of admissions statements showed score disparities between race, gender, and linguistic background (Breland et al., 1999), which were not fully explained by differences in writing proficiency.

Although blind-grading practices mitigate some of these biases, they do not eliminate them. One study of English as a Foreign Language (EFL) assessments found that even with blind grading and instructor training, human raters remained inconsistent, and their judgments varied systematically depending on the perceived linguistic and cultural identity of the student. In particular, the raters penalized deviations from native-like language use more harshly in students they presumed to be less proficient, even when the substantive content of the essays was otherwise strong (Schaefer, 2008).

Automated legal analysis scoring may be less susceptible to some of the kinds of errors and biases that affect human graders. This is almost certainly true for fatigue-related concerns: machines do not tire.[19] It is also possible—though less certain—that LLMs could produce evaluations that are less biased and more consistent. Indeed, as others have argued, reducing human grading biases "such as rater fatigue, rater's expertise, severity/leniency, scale shrinkage, stereotyping, Halo effect, rater drift, perception difference, and inconsistency" was one of the "key anticipated benefits" of using LLMs to grade essay-based exams (Kumar and Boulanger, 2020, at 2). Empirical evidence suggests that AES partially achieved these benefits, demonstrating greater consistency in some evaluation tasks than human scorers, including the reduction of interpersonal assessment bias (Attali and Burstein, 2006). Although our empirical design does not establish whether automated legal analysis scoring can achieve comparable reductions in human biases, the existing track record offers at least some reason for optimism.[20]

In weighing the relative strengths of AI and human grading, another question is whether errors by AIs and humans should be valued equally or whether mistakes made by AIs ought to be treated as more (or less) consequential. This debate parallels discussions about autonomous vehicles, where many argue that accidents caused by AI systems should carry greater weight than comparable accidents caused by human drivers (e.g., Krügel and Uhl, 2024; Geistfeld, 2017, at 1691-94). In this domain, some critics of autonomous vehicles point to mistakes made by

[19] In addition, any gradual decline in output quality across a chat session can be avoided simply by evaluating each answer with a separate query.

[20] Another approach instructors have used to address the difficulties with essay grading is to rely on multiple-choice exams (which can be computer graded without fear of inaccuracy). However, multiple-choice exams are only suitable for testing students' understanding of certain legal materials. Clients do not ask multiple choice questions. Essay questions better approximate the real-world experience of lawyers.

those vehicles as evidence for prohibiting them, while overlooking the proportionally larger number of equally catastrophic mistakes made by human drivers. The debate over automated legal analysis scoring might follow a similar path. Key issues in evaluating this question—though ones we cannot resolve here—include difficult ethical considerations. For instance, if machines produce the same aggregate level of mistakes as humans, does it matter whether they make *different* mistakes (e.g., assigning student A a lower grade and student B a higher grade, when a human grader would have erred in the opposite direction)? Similarly, if either the human or LLM tends to produce a large number of small errors, how should that be compared to an LLM or human counterpart that tends to produce a small number of larger errors?

5.2 Alternative Uses of Automated Legal Analysis Scoring

Although our results speak most directly to the prospect of replacing human grading of law school exams with AI grading, they also suggest alternative applications for automated legal analysis scoring, two of which we consider here. One possibility application is that automated legal analysis scoring could be used to *supplement* human grading. For example, divergences between LLM grading and human grading could be to identify outliers in human grading, which could be reviewed by the human grader to reduce the incidence of errors. Of course, instructors should consider the bias-inducing effects of *only* reviewing student answers for which there is a substantial difference between the human and LLM grading, but we argue that the potential value of the instructor-to-LLM grade discrepancies as a heuristic for human error is high.[21]

A second alternative application is to use automated legal analysis scoring to provide law students with individualized feedback on preliminary work product or mock exams. Evidence suggests that feedback can improve performance throughout law school, particularly for students who enter law school with lower academic indicators (Schwarcz and Farganis, 2017). But a well-known feature of legal education is that students, especially in their first year, often receive limited feedback on

[21] As an anecdotal example, results from Exam 1 showed a handful student answers that had more than a minimal discrepancy between instructor- and LLM-assigned grades. Upon review, most those discrepancies were arguably reflective of errors in the model's grading, but one of the discrepancies was clearly due to an error in the original, instructor-assigned grade. Had our approach been employed at the time of actual grading, that student's exam score could have been "corrected" (in this specific case, the student received a higher score than the grading rubric warranted).

their exam writing. Offering written feedback on mid-terms or practice legal analysis questions is time-consuming and costly for instructors, and thus likely to occur less than would be desirable for students.

Providing formative feedback in law school courses is not merely desirable for pedagogical reasons; it has also become a matter of compliance with American Bar Association accreditation standards. In early 2025, the ABA adopted revisions requiring that "all courses in the first one-third of the credit hours earned by students in the JD program include at least one formative assessment that allows students to evaluate their performance relative to the learning outcomes in the course." (American Bar Association, 2025) Predictably, many law professors have voiced concern that, depending on how this mandate is interpreted and implemented, it could impose substantial new burdens on already overextended faculty(Silverstein, 2024).

Our results suggest that automated legal analysis scoring could allow students to obtain reasonably accurate assessments of ungraded midterm exams or mock exam answers, and perhaps even meet new ABA requirements for formative assessments in the process. They also hint at the potential for AI feedback to extend beyond overall quality assessments, offering insights into the strengths and weaknesses of student work, and even suggesting concrete methods for improvement. The key question here—which we do not test in this paper but hope to examine in future research—is whether AI can generate accurate scores for individual elements within grading rubrics (as opposed to question-level scores), and do so reliably enough that those scores on individual elements can meaningfully guide students. An additional possibility, which likewise remains untested in this study, is that AI systems may be able to deliver helpful written feedback explaining why particular rubric scores were assigned and how students might improve their performance.

These potential methods for leveraging AI to improve law student writing could also extend beyond formal legal education into legal practice. As in law school, many junior lawyers receive limited feedback from senior colleagues on the quality of their work. Even when senior lawyers revise or rewrite junior lawyers' work, the latter are often left uncertain about the quality of their initial draft and forced to guess which edits reflect substantive deficiencies versus the senior lawyer's stylistic preferences. AI-generated feedback could both accelerate junior lawyers' learning and free senior lawyers' time. Although formal rubrics for evaluating the work of junior lawyers are not yet common, it is not difficult to envision law firms developing high-level criteria for certain types of tasks that could serve as the basis for such assessments.

A different implication of our results concerns the use of AI-based grading strategies in developing AI benchmarks. As noted earlier, as AI systems become increasingly capable of performing discrete legal tasks, it is critical to measure performance reliably across tools and over time. Although a number of legal benchmarks exist, a persistent challenge has been how to score AI outputs in response to legal questions. Some benchmarks attempt to avoid this problem by relying only on objective questions with unambiguous answers. But this approach necessarily narrows the scope of evaluation, as many important legal tasks do not have clear right and wrong answers (Fei et al., 2023). An alternative approach that has gained traction is to use AI itself to score outputs (Fan et al., 2025; ValsAI, 2025). However, because there is limited evidence on the reliability of such assessments, this strategy has been subject to criticism. Our results provide evidence that AI grading techniques may in fact be workable, particularly when AIs are supplied with structured rubrics. This suggests that building robust legal AI benchmarks may require not only curating a diverse set of legal analysis questions but also developing rubrics that can guide AI systems in evaluating responses.

6 Conclusion

Our analysis suggests that LLMs have the capacity to roughly approximate the grading of a law professor, with the best results coming from prompts that incorporate detailed rubrics. With LLMs' rapid improvement in text analysis tasks over just the last few years, we can probably expect continued improvement in the near future. Future analysis—including on different exams on different subjects—is likely to produce even greater agreement between humans and machines. Routine updates to this study will be necessary as the LLMs evolve.

Even as LLMs increasingly approximate human graders' performance, we acknowledge that logistical, institutional, and political challenges exist to replacing law professor grading. For instance, some law schools have longstanding rules requiring that professors (rather than, say, teaching assistants) personally grade exams and assign grades. It is unclear if machine-graded exams would comply with such rules. But even if they would or if the rules are changed, other political forces and path dependence may prevent law schools from replacing human graders with LLMs for the foreseeable future.

Even so, our findings suggest that automated legal analysis scoring could be effectively used for other valuable grading tasks in the very near future. For instance, professors could use LLMs as a supplement to their own grading, reviewing their own evaluations to detect any errors and bias. In addition, students might use LLMs and instructor-supplied rubrics to receive feedback on their practice exams, thereby potentially meeting otherwise burdensome ABA requirements to provide formative assessments in all first-year law school classes.

References

American Bar Association. 2025. Introduction to learning outcomes, assessment, and evaluation standards. Accessed: 2025-10-25.

Athey, S. and G.W. Imbens. 2019. Machine learning methods that economists should know about. *Annual Review of Economics 11*(1): 685–725 .

Attali, Y. and J. Burstein. 2006. Automated essay scoring with e-rater® v.2.0. *Journal of Technology, Learning, and Assessment 4*(3) .

Avery, J.J., P.S. Abril, and A. del Riego. 2023. ChatGPT, esq.: Recasting unauthorized practice of law in the era of generative AI. *Yale Journal of Law and Technology* 26: 64–127 .

Barocas, S. and A.D. Selbst. 2016. Big data's disparate impact. *California Law Review 104*(3): 671–732. https://doi.org/10.15779/Z38BG31 .

Blair-Stanek, A., P. Campbell, D.G. Gifford, G. Krishnamurthi, R.V. Percival, J. Sovern, M. Sweeney, D.B. Tobin, and L. Vertinsky. 2024. GPT gets its first law school b-pluses. SSRN Working Paper.

Blair-Stanek, A., A.M. Carstens, D.S. Goldberg, M. Graber, D.C. Gray, and M.L. Stearns. 2023. GPT-4's law school grades: Con law c, crim c-, law & econ c, partnership tax b, property b-, tax b. SSRN Working Paper.

Blair-Stanek, A., N. Holzenberger, et al. 2023. Can GPT-3 perform statutory reasoning? In *Proceedings of the 19th International Conference on Artificial Intelligence and Law (ICAIL 2023)*.

Bliss, J. 2024. Teaching law in the age of generative AI. Jurimetrics (forthcoming), SSRN Working Paper.

Breland, H., B. Bridgeman, and M. Fowles 1999. Writing assessment in admission to higher education: Review and framework. Technical report, The College Board. Reviews demographic bias in admissions writing assessments.

Büttner, M. and I. Habernal 2024. Answering legal questions from laymen in german civil law system. In *Proceedings of the 18th Conference of the European Chapter of the ACL (EACL 2024)*.

Carlson, D. and J.M. Montgomery. 2017. A pairwise comparison framework for fast, flexible, and reliable human coding of political texts. *American Political Science Review 111*(4): 835–843 .

Chai, F., J. Ma, Y. Wang, J. Zhu, and T. Han. 2024. Grading by AI makes me feel fairer? how different evaluators affect college students' perception of fairness. *Frontiers in Psychology* 15: 1221177. https://doi.org/10.3389/fpsyg.2024.1221177 .

Chilton, A., P. Joy, K. Rozema, and J. Thomas. 2024. Improving the signal quality of grades. *Journal of Law, Economics, and Organization 40*(3): 820–853. https:

//doi.org/10.1093/jleo/ewad012 .

Choi, J.H. 2025. Large language models are unreliable judges. SSRN Working Paper 5188865.

Choi, J.H., K.E. Hickman, A.B. Monahan, and D. Schwarcz. 2022. ChatGPT goes to law school. *Journal of Legal Education 71*(3): 387–400 .

Choi, J.H., A.B. Monahan, and D. Schwarcz. 2025. Lawyering in the age of artificial intelligence. *Minnesota Law Review 109*(1): 147– .

Choi, J.H. and D. Schwarcz. 2023. AI assistance in legal analysis: An empirical study. SSRN Working Paper.

Cyran, H. 2024. New rules for a new era: Regulating artificial intelligence in the legal field. *Journal of Law, Technology, & the Internet 15*(1) .

Engel, C. and R.H. McAdams. 2024. Asking GPT for the ordinary meaning of statutory terms. MPI Collective Goods Discussion Paper / SSRN.

Fan, Y., J. Ni, J. Merane, Y. Tian, Y. Hermstrüwer, Y. Huang, M. Akhtar, E. Salimbeni, F. Geering, O. Dreyer, D. Brunner, M. Leippold, M. Sachan, A. Stremitzer, C. Engel, E. Ash, and J. Niklaus. 2025. LEXam: Benchmarking legal reasoning on 340 law exams. arXiv preprint arXiv:2505.12864.

Fei, Z., X. Shen, D. Zhu, F. Zhou, Z. Han, S. Zhang, K. Chen, Z. Shen, and J. Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. arXiv preprint arXiv:2309.16289.

Frankenreiter, J. 2025. The other delaware effect. SSRN Working Paper.

Frankenreiter, J. and E. Talley. 2024. Sticky charters? the surprisingly tepid embrace of officer-protecting waivers in delaware. SSRN Working Paper.

Gamm, G. 2023. ChatGPT—what an attorney needs to know when using this new tool. Saint Louis University Law Journal Online.

Geistfeld, M.A. 2017. A roadmap for autonomous vehicles: State tort liability, automobile insurance, and federal safety regulation. *California Law Review 105*(6): 1611– .

Gray, M., J. Savelka, W. Oliver, et al. 2023. Can GPT alleviate the burden of annotation? In *Legal Knowledge and Information Systems*.

Guha, N., J. Nyarko, D.E. Ho, C. Ré, A. Chilton, A. Chohlas-Wood, A. Peters, B. Waldon, D. Rockmore, D. Zambrano, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *NeurIPS 2023 Datasets and Benchmarks Track*.

Hadi, M.U., R. Qureshi, A. Shah, M. Irfan, A. Zafar, M.B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili, et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints* .

Harasta, J., T. Novotná, and J. Savelka. 2024. It cannot be right if it was written by AI: On lawyers' preferences of documents perceived as authored by an LLM vs a human. *Artificial Intelligence and Law*. https://doi.org/10.1007/s10506-024-09422-w .

Hargreaves, S. 2023. "words are flowing out like endless rain into a paper cup": ChatGPT & law school assessments. *Legal Education Review 33*(1) .

Hoeijmakers, E.J.I., B. Martens, B.M.F. Hendriks, C. Mihl, R.L. Miclea, W.H. Backes, J.E. Wildberger, F.M. Zijta, H.A. Gietema, P.J. Nelemans, et al. 2024. How subjective CT image quality assessment becomes surprisingly reliable: Pairwise comparisons instead of likert scale. *European Radiology 34*(7): 4494–4503 .

Homoki, P. and Z. Ződi. 2024. Large language models and their possible uses in law. *Hungarian Journal of Legal Studies* .

Hunter, D.R. 2004. MM algorithms for generalized bradley-terry models. *The Annals of Statistics 32*(1): 384–406 .

Hussein, M.A., H. Hassan, and M. Nassef. 2019. Automated language essay scoring systems: A literature review. *PeerJ Computer Science* 5: e208 .

Juluru, K., H.H. Shih, K.N. Keshava Murthy, and P. Elnajjar. 2021. Bag-of-words technique in natural language processing: A primer for radiologists. *RadioGraphics 41*(5): 1420–1426 .

Katz, D.M., M.J. Bommarito II, S. Gao, P. Arredondo, J. Blackman, K.P. Bombardieri, et al. 2024. GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A 382*(2271): 20230254. https://doi.org/10.1098/rsta.2023.0254 .

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. 2018. Human decisions and machine predictions. *Quarterly Journal of Economics 133*(1): 237–293 .

Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer. 2015. Prediction policy problems. *American Economic Review 105*(5): 491–495 .

Kortemeyer, G. 2023. Toward AI grading of student problem solutions in introductory physics: A feasibility study. *Physical Review Physics Education Research 19*(2): 020163. https://doi.org/10.1103/PhysRevPhysEducRes.19.020163 .

Krügel, S. and M. Uhl. 2024. The risk ethics of autonomous vehicles: An empirical approach. *Scientific Reports 14*(1): 960. https://doi.org/10.1038/s41598-024-51313-2 .

Kumar, V. and D. Boulanger 2020. Explainable automated essay scoring: Deep learning really has pedagogical value. In *Frontiers in Education*, Volume 5, pp. 572367. Frontiers Media SA.

Li, T.W., S. Hsu, M. Fowler, Z. Zhang, C. Zilles, et al. 2023. Am i wrong, or is the autograder wrong? effects of AI grading mistakes on learning. In *Proceedings of the 2023 ACM Conference on International Computing Education Research.*

Liew, P.Y. and I.K.T. Tan 2024. On automated essay grading using large language models. In *Proceedings of the 2024 8th International Conference on Computer Science and Artificial Intelligence*, pp. 204–211.

Liu, Y., T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, et al. 2023. Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology*: 100017 .

Malouff, J.M., S.J. Stein, L.N. Bothma, K. Coulter, and A.J. Emmerton. 2014. Preventing halo bias in grading the work of university students. *Cogent Psychology 1*(1): 988937 .

Markoff, J. 2013. Essay-grading software offers professors a break. *The New York Times* April 5 .

Martin, L., N. Whitehouse, S. Yiu, L. Catterson, et al. 2024. Better call GPT: Comparing large language models against lawyers. arXiv preprint.

Martínez, E. 2024. Re-evaluating GPT-4's bar exam performance. *Artificial Intelligence and Law*. https://doi.org/10.1007/s10506-024-09396-9 .

Mitros, P., V. Paruchuri, J. Rogosic, et al. 2013. An integrated framework for the grading of freeform responses. In *The Sixth Conference of Learning International Networks Consortium (LINC).*

Mullainathan, S. and J. Spiess. 2017. Machine learning: An applied econometric approach. *Journal of Economic Perspectives 31*(2): 87–106 .

Murray, M.D. 2023. Artificial intelligence and the practice of law part 1: Lawyers must be professional and responsible supervisors of AI. SSRN Working Paper.

Nay, J.J., D. Karamardian, S.B. Lawsky, et al. 2024. Large language models as tax attorneys: A case study in legal capabilities emergence. *Philosophical Transactions of the Royal Society A*. https://doi.org/10.1098/rsta.2023.0159 .

Nitko, A.J. 1996. *Educational Assessment of Students.* Prentice Hall.

Page, E.B. 2003. Project essay grade: PEG, *Automated Essay Scoring: A Cross-Disciplinary Perspective.* Lawrence Erlbaum Associates.

Pierce, N. and S. Goutos. 2023. Why law firms must responsibly embrace generative AI. SSRN Working Paper.

Posner, E.A. and S. Saran. 2025. Judge AI: Assessing large language models in judicial decision-making. University of Chicago Coase-Sandor Institute for Law & Economics Research Paper No. 2503.

Savelka, J. and K.D. Ashley. 2023. The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. *Frontiers in Artificial Intelligence* 6: 1279794. https://doi.org/10.3389/frai.2023.1279794 .

Schaefer, E. 2008. Rater bias patterns in an EFL writing assessment. *Language Testing 25*(4): 465–493 .

Schwarcz, D. and D. Farganis. 2017. The impact of individualized feedback on law student performance. *Journal of Legal Education 67*(1): 139–167 .

Schwarcz, D., S. Manning, P.J. Barry, D.R. Cleveland, J.J. Prescott, and B. Rich. 2025. AI-powered lawyering: AI reasoning models, retrieval augmented generation, and the future of legal practice. Minnesota Legal Studies Research Paper No. 25-16; U. of Michigan Public Law Research Paper No. 24-058.

Silverstein, J.M. 2024. A critical perspective on formative assessment mandates. *University of Arkansas at Little Rock Law Review 47*(2): 189–243 .

Simon, D. 2023. More true confessions of a legal writing professor: Chat GPT makes a better therapist than a lawyer. Arizona Attorney (forthcoming 2023), Arizona Legal Studies Research Paper, SSRN.

Stollnitz, B. 2023. How GPT models work: For data scientists and ML engineers. Accessed: 2025-01-29.

Tiwari, A., P. Kalamkar, A. Banerjee, S. Karn, et al. 2024. Aalap: AI assistant for legal & paralegal functions in india. arXiv preprint.

Tomić, B.B., A.D. Kijevčanin, Z.V. Ševarac, et al. 2022. An AI-based approach for grading students' collaboration. *IEEE Transactions on Learning Technologies* .

Trozze, A., T. Davies, and B. Kleinberg. 2024. Large language models in cryptocurrency securities cases: Can a GPT model meaningfully assist lawyers? *Artificial Intelligence and Law*. https://doi.org/10.1007/s10506-024-09399-6 .

ValsAI. 2025. Legal AI report.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, Volume 30, pp. 5998–6008.

Yang, H. and K. Zhu 2022. Research on AI-assisted grading of math questions based on deep learning. In *International Conference on Internet of Things and Machine Learning (IOTML).*

Yang, R., J. Cao, Z. Wen, Y. Wu, X. He, et al. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *ACL Workshop/Proceedings.*

Appendix A: Answer Key (Redacted)

NOTE TO READERS: The below answer key has been redacted because the author of the exam may use some elements for an upcoming civil procedure course.

Summary of Answer for Question 2: For full points, students should make sure to address each of the following approaches to having the claims dismissed.

- Timing to Amend (3 points): [50-word explanation of the correct answer, an explanation of the application of the rule, and a note that courts are often liberal with amendments.]
- Joinder of Parties under Rule 20 (3 points): [40-word explanation of the correct rule to apply and the likely outcome.]
- Dismissal under 12(b)(2) (3 points): [50-word answer discussing the residency of the defendant, noting that the hypothetical creates a "borderline case" but also noting the most likely outcome.]
- Dismissal under 12(b)(3) (3 points): [75-word answer providing some detail regarding the application of the venue rules.]
- Dismissal under 12(b)(4) or (b)(5) (1 points): [20-word answer noting that there is not enough detail in the hypothetical to argue dismissal for improper service or process but that students should get a point for bringing it up as a possibility.]
- Dismissal under 12(b)(6) (2 points): [20-word answer noting that there is not enough detail in the hypothetical to argue dismissal for improper service or process but that students should get a point for bringing it up as a possibility, especially given how common 12(b)(6) motions are.]
- Removal to State Court (7 points total): [200-word answer noting that this is where the real meat of the question is. The answer covered the technical rules of removal in addition to general issues of subject matter jurisdiction, focusing specifically on diversity jurisdiction as the most fruitful avenue.]
- Discussion of Strategy (up to 3 points): [100-word answer noting that "Students can earn up to three more points by explicitly discussing strategy and plaintiff preferences independent of the conclusions they came to using the rules above," specifically the questions