Law Professors Prefer AI Over Peer Answers

Alejandro Salinas[1], Carly Frieders[1], Neel Guha[1], Sibo Ma[1],

Ralph Anzivino[2], Ian Ayres[3], Oren Bar-Gill[4], Omri Ben-Shahar[5], Stephen Friedman[6], George Geis[7], Sue Guan[8], Christoph Henkel[9], Stephanie Hoffer[10], Gregory Klass[11], Larasz Moody-Villarose[12], Sarath Sanga[3], Keith Sharfman[14], Justin Simard[15], Rebecca Stone[16], David Wishnick[11],

Julian Nyarko[1]

[1]Stanford University, [2]Marquette University, [3]Yale University, [4]New York University, [5]University of Chicago, [6]Widener University, [7]University of Virginia, [8]Santa Clara University, [9]Drake University, [10]Indiana University, [11]Georgetown University, [12]Purdue Global, [13]St. John's University, [14]Michigan State University, [15]University of California, Los Angeles

May 27, 2026

Abstract

Large language models (LLMs) are increasingly promoted as educational tutors, yet most evaluations focus on domains with a single ground truth. Many disciplines, however, hinge on judgment: reasoning, weighing ambiguity, and reaching defensible conclusions. Law provides a sharp test. We conducted a blinded evaluation of short-answer tutoring in contracts courses with sixteen U.S. law professors. Participants created 40 representative questions, wrote answers, and judged 2,918 anonymized comparisons between human and LLM responses. Professors rated LLMs far higher than their peers (average win rate = 75.33%), with models performing similarly to the best instructor. LLM responses were also rarely flagged as harmful (3.53%, vs 12.06% for professors). Preferences for LLM answers were consistent across evaluators and reflected shared professional standards. Our evaluation can be reliably extended to additional models by employing a separate LLM as a judge, rendering expert agreements an effective, scalable method to evaluate AI tutors in judgment-rich domains.

AI tutors have shown promise across a range of domains (1–9). In healthcare and STEM education, for instance, students often prefer—and sometimes learn more from—AI expla-

nations than from human ones (9). In physics courses, AI tutors can match or even outperform peer instruction (2). A common feature of these settings is that performance can be compared against a single correct, "ground truth" answer, rendering evaluation relatively straightforward (2; 9–11).

But many important domains are not like this. In fields such as history, philosophy, and law, correctness cannot be reduced to one solution (12). Instead, what matters is the exercise of judgment: weighing competing reasons, grappling with ambiguity, and reaching a defensible conclusion (13; 14). These are domains where two opposing answers can both be of high quality if each reflects underlying disciplinary standards rather than simple factual accuracy. Evaluating AI responses in such settings therefore requires a different approach—one that asks whether model outputs align with the latent professional standards that experts themselves endorse.

The legal domain provides a particularly clear test (14; 15). Some legal questions do admit unambiguous answers, such as whether a statute exists or what a case held. But the central pedagogical task in legal education is not just to convey facts, but to cultivate judgment. Students are trained to argue both sides of a novel case, and professors evaluate them by whether their analysis displays the qualities the discipline prizes. Prior studies of LLMs in law instead focused on contexts where answers can be marked as either right or wrong, or did not include a human baseline (16–19). While valuable, these approaches may not similarly capture whether AI responses can approximate the professional standards of judgment that legal instruction is designed to instill. Moreover, prior work has relied primarily on instruction-tuned models, without incorporating reasoning models. Yet reasoning capabilities—contextualizing governing principles, integrating them with novel facts, and deriving defensible conclusions—are central to legal problem-solving. This study is, to our knowledge, the first to systematically evaluate reasoning models in that pedagogical context.

We design an evaluation of AI-generated short answers that applies the profession's own internal standard of judgment. Figure 1 illustrates our design. Sixteen contracts professors from fourteen U.S. law schools—who all use the same casebook to teach the material—authored questions representative of those asked during office hours. From this pool we curated 40 representative questions spanning four instructional categories (Recall: Case or Code, Recall: Doctrine, Hypotheticals, Policy). Recall questions—whether relating to a case, code or doctrine—tend to be amenable to answers which can be evaluated against a ground truth, and where argumentative strength is of little importance. In contrast, hypotheticals present a short set of facts and ask how the law should be applied. Together with policy questions, which often center on legal or policy design under heterogeneous preferences, providing a strong answer in this category often relies on displaying careful

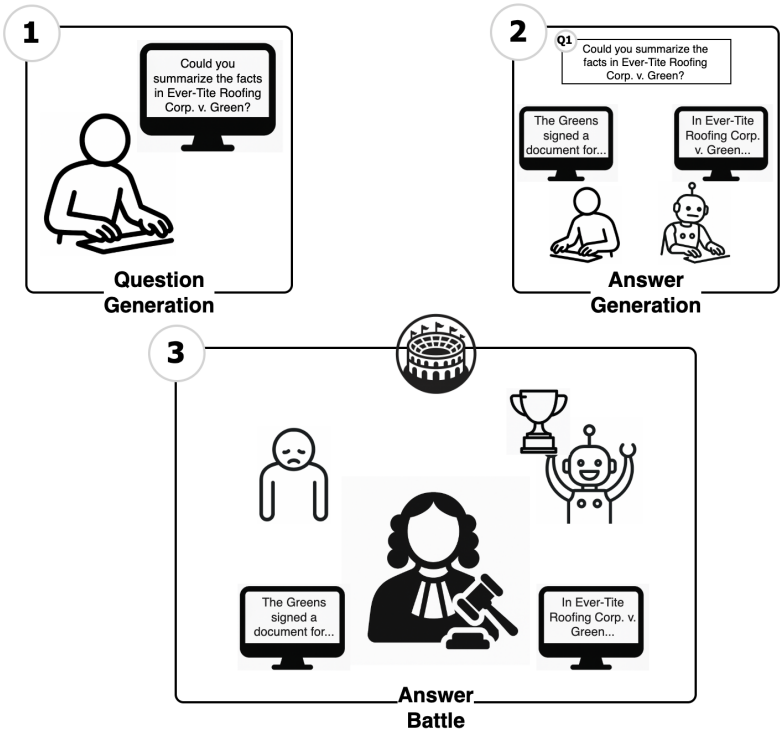


Figure 1: Overview of the evaluation process. In Stage 1 (Question Generation), instructors provided questions on contract law. In Stage 2 (Answer Generation), responses were produced either by those same instructors or by LLMs. In Stage 3 (Answer Battle), pairs of answers were blindly evaluated by instructors as judges, who indicated which response they would prefer to deliver to a student and whether any response was pedagogically harmful.

reasoning, weighing competing arguments and other latent, professional standards of quality—even if the relevant doctrine is now settled. In a second step, each professor wrote short answers to a subset of the 40 questions. To ensure that comparisons are course-agnostic, both the human and model answers were designed to avoid references to section-specific lecture remarks or slides. Similarly, the LLMs were not grounded in course transcripts. In a third step, we conducted blinded, forced-choice comparisons in which professors judged anonymized pairs of answers written either by their colleagues or by two LLMs. Among the different model families, we opted for Google's models because at the time, Google made explicit efforts to optimize their models for the educational context (20). Consequently, we included a stock version of Gemini 2.5 Pro and a retrieval-augmented NotebookLM with access to the casebook. Preference rankings have been shown to be a particularly effective method in ranking unstructured, open text responses, thus yielding advantages over more common, rubric-based evaluations especially where quality is a more elusive concept (21–26). To our knowledge, this study is the first to implement said methodology for evaluations in the educational sector (27–30). To probe whether any LLM advantage might be driven by surface-level writing style rather than substantive content, we additionally engineered a set of lexico-syntactic features—answer length, structural organization, reasoning nuance, legal anchors, confidence tone, clarity, and pedagogical support—and tested how much of the preference pattern they could explain. Each professor completed approximately 150–200 pairwise evaluations, selected the better answer, and could flag any answer as pedagogically "harmful."

We present four main findings. First, LLMs meet—and often exceed—the professional standard as defined by expert preference. Gemini 2.5 Pro outperformed all but one instructor in head-to-head comparisons (average win rate against all instructors = 75.92%), though the difference between Gemini and the better-ranked instructor was not statistically significant. NotebookLM, by contrast, outperformed every human instructor, with one tie (average win rate = 74.75%). Second, the LLM advantage was similar across all category questions. Third, harmfulness rates for LLMs were low (Gemini 3.41%, NotebookLM 3.64%), compared to the wider dispersion among professors (1.00–39.75%), underscoring that the risk of pedagogically problematic responses is comparable to that of the best human instructors. When evaluating peer-written answers, each professor on average preferred LLM responses over responses generated by human instructors, suggesting that model outputs were not merely appealing to a particular subset of evaluators. Fourth, the engineered textual features explain only part of the LLM advantage: in calibration analyses, observed LLM win rates systematically exceed the win rates predicted from lexico-syntactic differences alone, indicating that the preference for LLM answers is not reducible to length, clarity, or other stylistic markers.

To assess whether this performance reflects recovery of a shared professional standard rather than idiosyncratic preferences, we measured agreement among professors on overlapping trials. Observed agreement exceeded the level expected if judgments were entirely idiosyncratic, indicating that the LLMs' success reflects alignment with common disciplinary criteria. Complementary analyses using rescaled inter-coder correlation lend additional support to that hypothesis.

Because expert judgment is costly, our human evaluation was necessarily limited to two model variants (Gemini 2.5 Pro and NotebookLM). These models were frontier models when the human evaluation was conducted (August 2025). To extend the comparison to a broader set of systems–including models released after our human evaluation–we employ an "LLM-as-judge" framework, an increasingly common methodology for scaling evaluation of open-ended generation (31–33). We note that LLM judges have documented limitations—including position bias, verbosity bias, and a tendency to favor outputs from models in their own family (34; 35)—and we therefore first validate our chosen judge (Llama-4 Maverick) against the leave-one-out majority of human evaluators before deploying it. We then use it to rank Claude Opus 4.7, ChatGPT 5.4, Gemini 3.1 Pro, Claude Opus 4.1, ChatGPT 5, Gemini 2.5 Pro, Gemini 2.5 Flash (with and without thinking budget), Gemini 2.0 Flash, NotebookLM, and a commercial AI tutor. All models outperform human instructors. Claude Opus 4.7 ranks highest, followed by the other newest-generation models, illustrating that the gap between the strongest LLMs and expert instructors has continued to widen since our human evaluation was conducted.

Taken together, our findings suggest that LLMs can successfully capture latent professional standards in domains without ground-truth answers, at least in the constrained setting of short, office-hours-style responses. More broadly, we offer a method for evaluating AI tutors in such settings: inter-evaluator agreement—typically used as a reliability check—can instead serve as the basis for interpreting whether AI performance reflects mere personal preference or genuine alignment with a shared standard of judgment.

Results

Sixteen U.S. contract law professors participated both as instructors and as judges. As instructors, they answered questions typically asked by first-year students in office hours. In parallel, two LLMs (Gemini 2.5 Pro; NotebookLM grounded in the relevant casebook) answered these questions. As judges, the professors then completed 2,918 blinded, forced-choice comparisons (median per judge: 200), each time indicating which of the two anonymized responses, from the instructor or the LLM, they would rather give to a student. Our primary

metric is the *win-rate*—the fraction of pairwise trials in which a model's or instructor's response was preferred. Secondary metrics are (i) each judge's *LLM-preference rate* in human-vs-LLM trials (a baseline tendency to pick the LLM) and (ii) the *harmfulness rate*, the share of answers flagged by judges as likely to mislead or hinder learning. Full computational details and uncertainty estimates appear in Supplementary Information B.

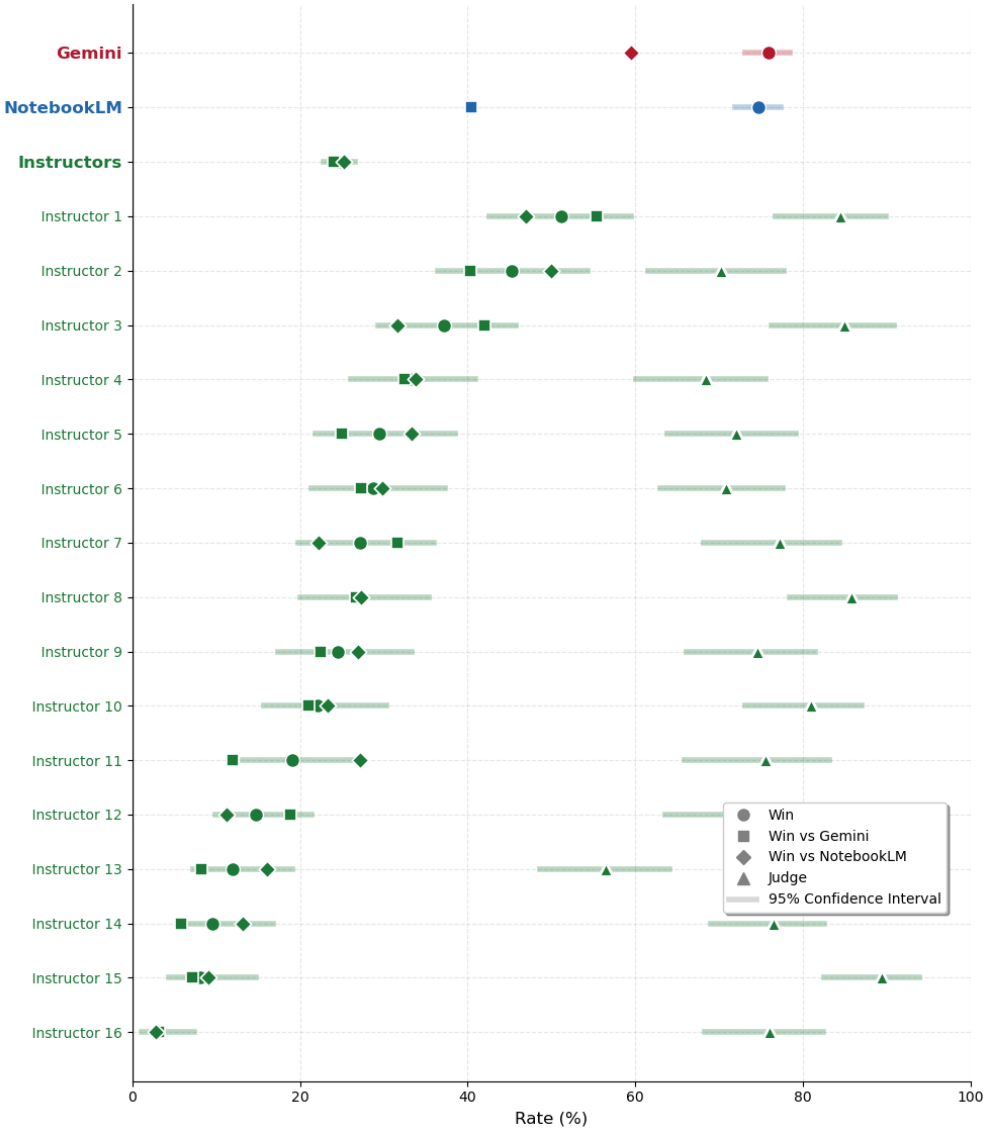


Figure 2: Win rates of instructors compared to LLMs (Gemini and NotebookLM). Circles indicate win rates with 95% Wilson score confidence intervals; diamonds and squares represent win rates against NotebookLM and Gemini, respectively. Triangles indicate LLM-preference rates at the judge level. Models are shown at the top, followed by individual instructors.

LLM answers are preferred over instructor answers. Rather than grading correctness, judges selected the response they would rather have a student be exposed to. When considering the average win rate, LLMs far outperformed human instructors, with Gemini achieving 75.92% across all trials against instructors, while NotebookLM won 74.75%, as shown in Fig. 2. Indeed, the LLMs were on par with the best human instructors who participated in the study, with NotebookLM outperforming every participant with one tie. Individual instructors' average win rates against both LLMs spanned 2.96-51.15%, with the pooled instructor average at 24.67%. Results are substantially equivalent when removing answers to questions that instructors indicated limited familiarity with (Supplementary Information C). However, individual win rates alone do not make efficient use of information in the pairwise preference ranks. Instead, it is common practice to analyze the inherent strength through Bradley-Terry models (36). A Bradley-Terry model pools information across all pairwise comparisons to infer a global ranking on a common scale, thus allowing inference about participants even from pairs in which they do not directly face each other. Doing so in this case reveals that Gemini is the best-performing model, and its difference with NotebookLM is statistically significant, as shown in Supplementary Information D. The best-performing human instructor was ranked second, while NotebookLM was ranked third. When considering professors as judges, the median LLM-preference rate in human-vs-LLM trials was 75.81%. All judges preferred LLMs over humans, with a minimum win rate at the judge-level of 56% in favor of LLM-generated answers.

The LLM advantage persists across question-type categories. To contextualize domain variation, we group questions into one of four types: *Recall-Case/Code*, *Recall-Doctrine*, *Hypotheticals* and *Policy* (see the *Methods* section for definitions). Figure E.1a in Supplementary Information E.1 disaggregates win rates by category. The LLM advantage appears in each category, ranging from 74.24% for Hypotheticals to 77.17% for Recall-Case/Code (Gemini), and from 72.69% for Hypotheticals to 76.80% for Recall-Case/Code (NotebookLM). In Supplementary Information E.2, we further disaggregate responses according to whether the question has a *clear* answer, is based on a *false premise*, is contained *in the casebook*, lacks a clear answer due to the *absence of a clear rule/precedent* or because it requires the application of a *vague standard*. Since instructors provided these labels, the number of judgments across them is imbalanced and we present results using a win rate model only (Figure E.1b). They show that LLMs consistently outperform human instructors, with no significant differences across labels. When comparing Gemini to NotebookLM, it is striking that there does not seem to be a meaningful advantage of the latter for questions with an answer contained in the casebook (42.07% win rate vs Gemini, compared to 40.36%

in the broader set). This is despite the fact that NotebookLM's responses are grounded in the casebook through RAG. Together, the results indicate that LLMs consistently provide highly rated answers even when quality is tethered to an implicit, professional standard rather than to factual accuracy.

Models rarely give harmful answers. Results vary for humans. The rate of harmful answers captures how often judges identified a response as likely to hinder student learning. Aggregating across responses, Gemini's harmful rate was 3.41% and NotebookLM's was 3.64% (LLMs pooled: 3.53%; 95% Confidence Interval: 2.44-4.61). Instructor responses encompass a broader range of 1.00-39.75% with an average harmfulness rate of 12.06% (95% Confidence Interval: 9.02-15.10), significantly higher than the models' ($p = 4.7 \times 10^{-7}$). Fig. F.1 display these rates; aggregation details and interval construction are provided in *Methods* and Supplementary Information F.

Judge preferences converge beyond a mere preference for LLMs, indicating a shared professional standard. Next, we assess the extent to which preferences of our human judges converge. Focusing only on comparisons of Gemini to human instructors where at least two judges saw the same answer pair, we contrast the observed intercoder agreement for each question to the smallest agreement rate that is consistent with observed LLM win rates. This minimum represents the expected intercoder agreement if judges, aside from a shared preference for LLMs, were merely following their own private leanings (details and the formal definition of this lower bound appear in Supplementary Information G). Figure G.1 shows that, across questions, observed intercoder agreement tends to be substantially higher than the LLM preference necessitates. Convergence is visible in every category, and is strongest for *Policy* questions (observed mean ≈ 0.77), followed by *Recall—Case/Code* (≈ 0.68), with *Hypotheticals* and *Recall—Doctrine* close behind (both ≈ 0.65). In short, when judges chose between LLM and instructor answers, they tended to apply a tacit, but common, evaluation rubric—rather than idiosyncratic tastes. This phenomenon is pronounced even in settings without clearly defined, ground truth answers—consistent with the presence of a shared professional standard.

All judges are similarly sensitive to answer quality. Figure 2 suggests that all judges prefer LLM responses over human instructor responses. However, it is possible that this aggregation masks meaningful heterogeneity. In particular, it appears plausible that LLMs give "good enough" answers that appeal to the median instructor, but that particularly talented instructors are able to identify significant flaws in their responses. To investigate

the relationship between the quality of the judge and their sensitivity to answer quality, we plot the quality of the human instructor answer a judge sees against the quality of the judge, and assess the judge's propensity to pick the instructor answer as a function of these two factors. We proxy judge quality by the average win rate of the answers they provided as instructors. To estimate the strength of human instructor answers seen by a judge, we compute the instructor answer's win rate, excluding the judge's own vote ("Leave-One-Judge-Out (LOJO) Win Rate"). Figure H.2 in Supplementary Information H displays our findings. Unsurprisingly, for judges of any quality, the probability of preferring the human instructor answer increases with the quality of the answer itself. But beyond that, we do not detect significant trends across the horizontal axis. This assessment finds further support in Figure H.1, which directly assesses the correlation along the horizontal axis. In effect, our results indicate that judges of all quality tend to exhibit similar patterns in their preference for instructor answers, lending support to the hypothesis that LLM responses have broad, general appeal. Additional details on the construction of these metrics are provided in Supplementary Information H.

Textual features only partially account for the LLM advantage. Is the LLM advantage plausibly driven by *how* the answer is written or by *what* the content of the answer is? To investigate this question, we generate several textual features of answer quality informed by best practices in the AI tutor and AI education literature (20; 27; 29; 37). In particular, we measure an answer's *structural organization*, *reasoning nuance*, *legal anchors*, *confidence tone*, *clarity*, *pedagogical support*, and *length*. All of these features are based on lexical/syntactic information and do not take into consideration semantic content. Figure I.1 in Supplementary Information I shows that only answer length is significantly positively correlated with a higher win rate, whereas *clarity* (Flesch reading-ease score (38)) and *pedagogical support* (the sum of the rate of question marks and the scaffolding rate per answer (39; 40)) are negatively correlated with answer win rate. We note, however, that none of these estimates can be understood causally absent strong assumptions. More importantly, Figure 3 compares observed LLM win rates against predicted win rates estimated purely from these textual features and shows that LLMs consistently perform better than expected, based on their lexical/syntactic features only. The difference is particularly pronounced for the decile with the lowest predicted win rate. These results suggest that the LLM advantage cannot be fully explained by common textual features, lending support to the hypothesis that the answer quality is at least partly driven by the content of the answer.

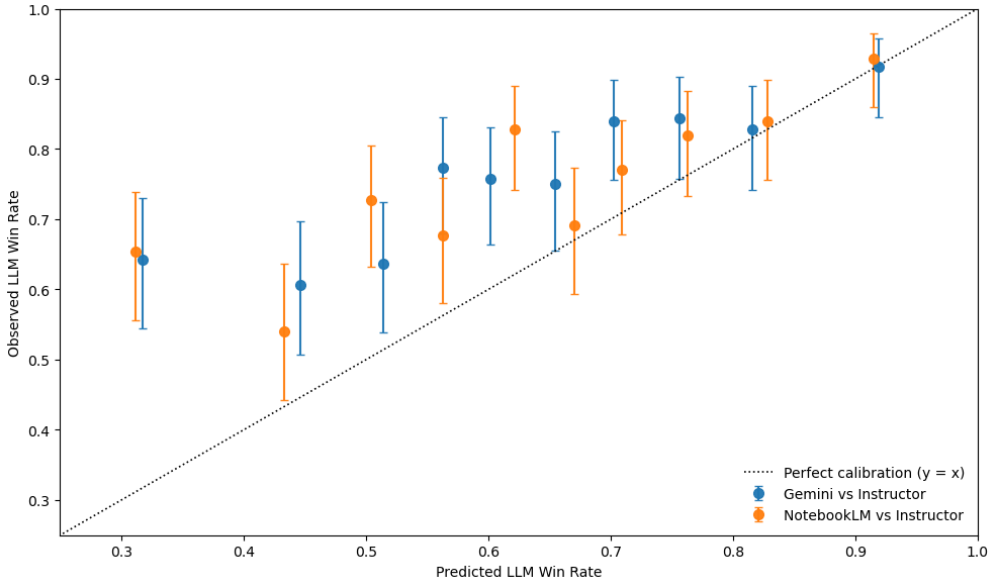


Figure 3: Calibration of model-predicted win probabilities vs. observed win rates (LLM vs. Instructors). Points show the observed win rate in each decile of the predicted probability (horizontal axis uses the bin mean predicted win rate). Vertical bars are Wilson 95% confidence intervals. The dotted line ($y = x$) indicates perfect calibration. Blue: Gemini vs. Instructor; orange: NotebookLM vs. Instructor. Predicted probabilities derived from a logit model regressing preferences on Δ-features, i.e. the difference in responses with respect to legal anchors, reasoning nuance, structural organization, confident tone, clarity, length, and pedagogical support. Systematic elevation of observed points above the diagonal implies that the included textual features do not fully explain the LLM advantage.

Ranking additional models via an "LLM-as-judge" framework. Our participants were constrained in time and thus could only provide a limited number of judgments, cabining our main analysis to two LLMs (Gemini 2.5 Pro and NotebookLM). To scale the evaluation across a broader set of systems, we examined whether an LLM can stand in as a judge ("LLM-as-judge" framework). To that end, we assess the capabilities of *Llama-4 Maverick* as a judge and find that it is able to reliably recover the majority vote of our human judges. Indeed, in Figure J.1, we show that it is about as reliable as the most aligned judge participating in our study (for details, see Supplementary Information J). We therefore used *Maverick* as a judge to scale our evaluation across nine additional models, including a model employed by a commercial online tutoring platform running on Gemini 2.5 Pro. Together with *NotebookLM*, the commercial AI tutor is the only model grounded in the casebook. Results are displayed in Figure 4. *Claude Opus 4.7* ranked highest, followed by *ChatGPT 5.4* and *Gemini 2.5 Pro*; *ChatGPT 5*, *Claude Opus 4.1*, and *Gemini 3.1 Pro* cluster next, with human instructors ranked lowest. Every AI model evaluated outperformed human instructors, on average. Three patterns stand out. First, *Gemini 2.5 Flash Thinking* was rated significantly stronger

than *Gemini 2.5 Flash No Thinking* (without a thinking budget), illustrating the importance of reasoning capabilities in answering legal questions. Second, last generation's *Gemini 2.0 Flash* ranked at the bottom of the AI models—broadly consistent with rapid progress in capabilities over time. That said, the relationship between model recency and performance is not monotonic: *Gemini 3.1 Pro* ranks below *Gemini 2.5 Pro* in our setting. We hypothesize that this result may stem from the apparent discontinuation of explicit efforts in improving Google's models for learning environments after the release of *Gemini 2.5 Pro*. Third, both the *Commercial AI Tutor* and *NotebookLM* rank well below the stock *Gemini 2.5 Pro* they are built on, despite being grounded in the casebook.

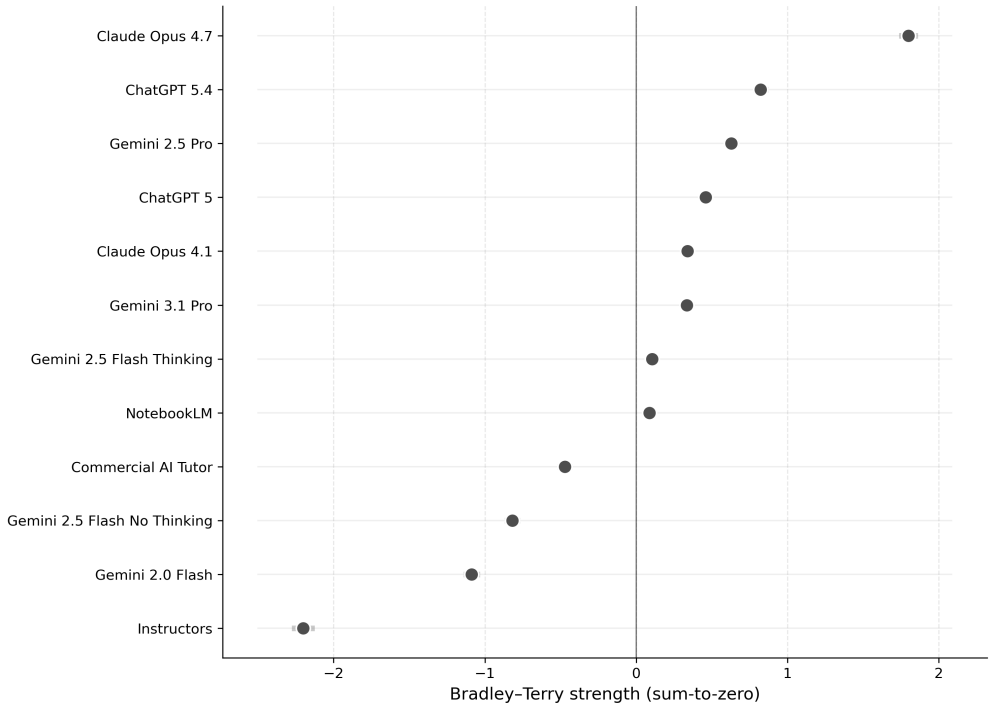


Figure 4: Bradley-Terry ranking of models with 95% confidence intervals, model strength summing to zero. Judge: *Llama-4 Maverick*. Each point shows the estimated latent strength; horizontal bars show maximum likelihood estimation confidence intervals from the observed Fisher information.

Discussion

We find that contemporary LLMs can already provide short-answer tutoring in first-year Contracts courses that is rated as comparable in quality to those given by the strongest instructors who participated in this study. This is despite the fact that high quality answers often require surfacing a shared, professional standard rather than merely matching a fixed

rubric. In blind, head-to-head judgments, instructors consistently preferred the LLM answer. Conditioning on response length and other surface markers shows that textual features do not fully explain these preferences; the remaining advantage is plausibly driven by substantive content quality rather than stylistic markers alone. Instructors also detected harmfulness at higher non-trivial rates for human answers as opposed to LLMs'. Together, these results suggest that in domains like law, where "quality" is partly a matter of judgment, LLMs can converge on norms multiple experts recognize as better answers–even when no explicit grading rubric is provided.

In our model setup, we intentionally opted for default, platform-level settings to simulate how students would likely use LLMs in practice; beyond a standard prompt asking the model to act like an instructor and a target length range (derived from faculty example answers), we did not tune decoding parameters or temperature. This design choice renders our estimates realistic, but plausibly conservative: we show strong performance without optimization, implying that priors useful in legal pedagogy are already encoded in the base models. In our blinded expert evaluation, Gemini 2.5 Pro (stock-version) outperformed NotebookLM, both under the win-rate outcome and in pooled Bradley-Terry rankings, despite NotebookLM being grounded in the casebook. In our complementary LLM-as-judge analysis, this same pattern extends to a commercial AI tutoring product that is also built on Gemini 2.5 Pro and grounded in the casebook. We interpret LLM-as-judge results with appropriate caution, as the methodology is known to exhibit biases such as those involving position, verbosity, and in-family preference (34; 35); here, however, the cross-developer robustness of the LLM advantage and our validation of *Llama-4 Maverick* against the human leave-one-out majority give us reasonable confidence in the broader pattern, if not in the precise ordering of closely-ranked systems.

To us, it is unexpected that a stock model would outperform RAG-grounded variants, particularly those optimized for an educational context. Although we cannot verify the training data composition, we propose three non-exclusive explanations. First, if the base model already possesses strong prior knowledge and adequately encodes the doctrinal structure for first-year Contracts, retrieval may add little marginal benefit. Second, overloading the context with long documents can dilute relevant material and introduce noise. This is a well-documented phenomenon–sometimes called the "lost in the middle" or "distracting effect"–where LLMs underutilize mid-context information or misweight weakly relevant paragraphs. Performance tends to degrade further when inputs approach very long context windows (e.g., 64k tokens). In our setting, NotebookLM and the commercial AI tutor were grounded in the full casebook (divided into chapters) rather than highly optimized chunks of text, making context saturation and retrieval noise plausible contributors to the observed

outcomes (41–44). At the same time, we note that optimized, user-driven chunking would not reflect a realistic use. Third, it is possible that additional platform layers like custom system prompts could inadvertently counteract some of the pedagogically helpful features of Gemini 2.5 Pro, chief among them LearnLM. We did not–and could not–control internal toggles. We emphasize that these are hypotheses generated by our design rather than definitive causal claims. They motivate follow-up experiments that manipulate retrieval scope, chunking strategies, and optimization pipelines to better measure when RAG and well-intended educational adjustments enhance legal tutoring performance and when they hinder it, both in our domain and in the broader literature.

While LLM responses are generally preferred over those of human instructors, our evaluation setting does not allow us to directly measure the extent to which instructor preferences are satisfied. It is at least theoretically possible that LLMs, although generally delivering stronger responses, still generate answers that are merely viewed as "good enough." This would especially be the case if instructors had idiosyncratic preferences over teaching that are satisfied neither by other instructors nor LLMs. While our results do not lend support to this hypothesis, we note that our setting does not allow for a direct test. At the same time, in real world deployments, LLMs can be grounded in idiosyncratic, course-specific content, such as a lecture transcript or notes, offering a practical way to address heterogeneous preferences of that kind. A separate scope question concerns the sample of instructors itself. Of the sixty professors who taught from the relevant casebook in the four years preceding the study, sixteen consented to participate. Comparing respondents to non-respondents on a set of observable characteristics, the resulting sample over-represents professors at T14 law schools (38% of respondents vs. 22% of the full pool) and, more modestly, under-represents female professors (25% vs. 33%). Respondents are also more likely to be tenured (94% vs. 83%), though their median time since tenure is somewhat shorter (16 vs. 19 years). We report these imbalances for transparency, the "shared professional standard" we estimate is the one held by this particular subset of contracts instructors, and the extent to which it generalizes to the broader population of legal educators is an empirical question we cannot resolve here. Table K.1 in Supplementary Information K reports the full distribution of these characteristics for respondents and non-respondents.

Our findings complement—and in important respects diverge from—recent evaluations of AI in legal and related domains Ouellette et al. (16); Dahl et al. (45); Magesh et al. (46); Choi (47); Shojaee et al. (48); Petrov et al. (49), many of which conclude that leading models are not yet acceptable for engaging in legal analysis or for providing student advice. We read these results as task- and evaluation-dependent rather than a global and definitive consensus. Three distinctions help reconcile patterns across studies. First, our outcome is

an expert preference between two plausible answers under blinding, rather than accuracy against a single, prespecified rubric. That choice is deliberate: in many respects of legal pedagogy, "quality" reflects a shared but implicit professional standard rather than one canonical solution. Second, our paired-comparison design anchors LLM performance to a human baseline produced by the same instructor population, enabling direct comparisons on questions actually relevant to their courses. Third, to our knowledge, our study is the first in this space to explicitly evaluate reasoning-oriented models. Providing adequate legal answers requires more than pure exposition; it often demands a substantial degree of legal reasoning (50; 51). This is especially true for hypotheticals, where established rules must be applied to a novel fact pattern in order to derive a new conclusion. Because such hypotheticals are variations on–but typically not directly represented in–the training data, high-quality answers cannot be produced merely by retrieving memorized responses. Instead, LLMs must be able to contextualize the governing legal principles, integrate them with the facts at hand, and work through their implications to reach a defensible conclusion–a task reasoning models are particularly suited for.

Limited instructor availability in legal education is often rooted in capacity constraints and professional demands, not a lack of willingness. Our results foreground a novel benefit for legal education: an always-available, short-answer clarification channel. Where students currently rely on peers or sporadic email exchanges with faculty, a reasonably reliable AI clarification channel that reflects shared professional norms could, in principle, be accessible on demand. In our blinded experiment, judges did not systematically prefer human-authored over model-generated answers, suggesting that carefully constrained AI assistance does not degrade perceived answer quality. To support reliability, implementations should establish clear scope limits, refusal policies when uncertain, deterministic decoding, visible citations to course materials, and straightforward pathways to escalate further questions and edge cases to instructors.

Our design evaluates answer quality—which response an expert would prefer to deliver under blinding—rather than learning impact. We therefore treat our results as an encouraging first indication that LLMs can reflect a shared professional standard in short-answer Contracts pedagogy—not as proof of improved student outcomes, and not as evidence regarding richer tutoring interactions. The next steps could focus on the empirical and design choices of deployment. On the empirical side, we suggest course-embedded randomized controlled trials (RCTs) that randomize structured AI support (with logging) and evaluate using blinded methods (following Kestin et al. (2)). On the design side, an error-aware tutor would be ideal: refuse when premises look ill-posed; constrain scope to the casebook/-syllabus; ground responses in citations; and use deterministic decoding with explicit brevity

limits (since, as shown, length does not fully explain the LLM advantage). Retrieval implementation should include test chunking strategies and retrieval thresholds to minimize context dilution and distraction, and report when RAG helps versus when the base model suffices. Because our primary outcome is pairwise expert preference, departments can also fit lightweight preference models to nudge tutors toward a local standard of judgment (per course, professor, unit) without exposing student data–conceptually analogous to objectives in direct preference optimization / reinforcement learning from human feedback (22; 24; 25), but grounded in faculty choices about what best supports their students. In addition, randomization in answer styles could be used to further investigate our finding that some textual features thought of as pedagogically useful (like scaffolding) do not seem to predict higher quality ratings.

At the same time, emerging work suggests that LLM use may influence learning outcomes in unexpected ways. For example, Melumad and Yun (52) find that exposure to LLM-generated summaries can have a small but significant negative effect on retention compared to traditional internet searches. In the legal domain, however, internet searches are unlikely to demonstrate an appropriate baseline. High-quality legal answers depend on articulating a shared professional standard, yet existing search engines rarely aim to provide such responses. In particular, open-ended web searches typically do not yield much direct guidance that extends beyond mere exposition–that is, unless the user is already an expert. Despite these differences, we emphasize that further research is needed to rigorously assess the effects of LLM use on learning outcomes.

Methods

Background

Law students in the U.S. have significant autonomy in how they design their curriculum. However, at many law schools, the first year is dedicated to a set of classes that are not only designed to teach students fundamental skills in legal reasoning and advocacy, but also cover topics that appear on most bar exams. One of such courses is Contract Law (or Contracts), which covers the law of voluntary exchanges and risk allocations, such as sales, insurance or lease agreements. Because Contracts is usually a mandatory class, there is no self-selection of students into the course, making it a particularly appropriate subdomain to study. When addressing questions in contract law, instructors confront different types of inquiries, each varying in the extent to which they admit of a ground-truth answer. To categorize these questions, we employ the following taxonomy:

- **Recall–Case/Code**: The question primarily asks for a summary of facts or law stated in a particular case, statute or authoritative source of contract law.
- **Recall–Doctrine**: The question asks for an explanation of one or more doctrinal concepts.
- **Hypotheticals**: The question requires analysis of how the doctrine would apply to a novel set of facts rather than a mere restatement of the relevant rules.
- **Policy**: The question asks about policy rationales or implications of a legal rule.

Although there is some discretion as to what aspects of a case or doctrine to emphasize, answers to *Recall* questions lend themselves to determinations of accuracy. At the same time, they are generally considered straightforward, as an accurate answer primarily rests on memorization. In contrast, *Hypotheticals* explore a previously unseen set of facts. A good answer requires the transfer of learned rules to new contexts and the ability to draw out parallels and differences to previously seen case, a task commonly referred to as "legal reasoning" (50; 51). Hypotheticals are a core instrument in legal education, as they allow the instructors and students to explore the boundaries of doctrines and statutes for plausibility. They often make up the largest proportion of questions on law school exams. At the same time, hypotheticals often do not admit of a single correct answer. Different responses can be equally strong if they display careful reasoning, grapple with ambiguities, and reach defensible conclusions. This makes hypotheticals a paradigmatic example of tasks where evaluation depends not only on ground truth, but also on whether an answer reflects the latent professional standards that define sound legal judgment. Similarly, policy questions lack determinate answers: they invite students to weigh competing considerations, articulate tradeoffs, and defend normative positions. What distinguishes a strong response is not only correctness, but also whether it reflects the discipline's standards for framing arguments and exercising judgment. In addition to these labels, instructors who submitted the questions had the opportunity to select one of the following labels:

- **Clear Answer**: There is one clear answer or narrow set of non-conflicting possible answers to a question.
- **Vague Standard**: The answer is unclear because the question raises an issue which is not obviously resolved by the relevant legal standard.
- **Absence of Clear Rule/Precedent**: The answer is unclear because there is no or conflicting authority on an issue necessary to resolve the question.

- **In Casebook**: The casebook (Ayres, Klass & Stone's Studies in Contract Law) provides sufficient basis for answering the question, though inference may be required and the answer may not be complete.

- **False Premise**: The question reflects a clear misunderstanding of some aspect of contract law which needs to be corrected.

Questions with *clear* answers admit a single correct response, whereas those employing a *vague standard* or have *absence of a clear rule or precedent* rest only on the strength and thoroughness of the argumentation, irrespective of which result is achieved. For questions with an *answer in the casebook*, RAG might be of particular importance. Questions with a *false premise*, it may be hypothesized, pose particular challenges for LLMs, which often shy away from disagreeing with the user (53).

Design

We ran a blinded, head-to-head evaluation of short-answer tutoring in first-year Contracts using a national sample of instructors who teach from the same casebook (54). All sixty professors who used the casebook to teach the class in the past four years were invited to participate. Sixteen of them, representing fourteen U.S. law schools, consented and completed all phases under a Stanford IRB Exempt, Non-Medical determination (Protocol #80628). To assess the representativeness of the participating sample, we collected publicly available characteristics on respondents and non-respondents along several dimensions, including gender, law school ranking, tenure status, and years since tenure. The respondents overrepresent professors at T14 law schools (38% of respondents vs. 22% of the full pool) and, more modestly, under-represent female professors (25% vs. 33%); respondents are also more likely to be tenured (94% vs. 83%), with a median time since tenure of 16 years compared to 19 in the full pool. Table K.1 in Supplementary Information K presents the full comparison. Concretely, instructors first authored questions from the shared casebook and wrote brief answers, while we generated matched LLM responses. We then anonymized and lightly standardized all text, randomized pairings and left-right position, and, lastly, asked instructors to choose which of two anonymized answers they would rather deliver to a student. Figure 1 illustrates our pipeline, which is further described below. Email instructions can be found in A.3.

In the first stage, instructors submitted a total of eight questions, two for each of the four categories in our taxonomy. The submitted questions were intended to reflect inquiries students make after class or during office hours, thus mimicking those that might be directed

at a tutor. We also asked for them to span subject matter and difficulty while avoiding niche topics. Instructors further had the opportunity to indicate whether they viewed their question as admitting a clear answer, whether the answer was in the casebook and whether the question was based on a false premise. Each question included a model answer to signal scope and pedagogical intent; these were not included in the evaluation subset, but were later used to guide the answer length of the LLMs (see below). From this pool of questions, the research team curated forty representative questions (ten per category). Table A.1 in Supplementary Information A.2 presents all selected questions from each category. In a second stage, each instructor answered ten questions they did not author, simulating how they would respond in an office-hours setting. Instructions emphasized brevity (targeting ≤ 3 minutes of writing), could allow for uncertainty, and–to simulate realistic circumstances– instructors were asked not to conduct additional research before answering. That said, they could indicate if they were not familiar with the topic of a question and if they had to conduct research to produce the answer.

We generated model answers to the same forty questions using LLMs. Because assessments via pairwise preference ranks are relatively costly, rather than creating a holistic evaluation, we opted to focus our human evaluation on a small number of the most promising models, while later extending the analysis through an "LLM-as-judge" framework to additional LLMs. Among the different developers, we chose LLM responses from Google's Gemini family. This is because, through LearnLM, Google has documented concrete efforts to optimize its models for an educational context. We thus hypothesized that the Gemini family would likely yield the strongest results in our context.

The first model we used to generate answers is a stock version of Gemini 2.5 Pro, which integrates the LearnLM educational tuning in its architecture (20; 55). The model was prompted to generate an answer in the context of office hours without external materials. In addition, we generated answers via Google's NotebookLM (56). While NotebookLM largely relies on Gemini 2.5 Pro, unlike the stock version, NotebookLM employs retrieval-augmented generation to ground its answer in provided reference material. We used the relevant casebook–shared by all instructors–as a reference source. Both models were further guided to behave like law professors through a standardized system prompt. Based on the model answers provided in the first step by instructors, we further guided the LLMs to limit the length of their answers to those given by humans (50-108 words, ideally $\tilde{9}0$; up to 155 words if nuance required) [1]. The complete system prompts for both models can be found

[1]During the first stage of the study, instructors were asked not only to propose representative questions that students might ask after class, but also to supply the answers they themselves would typically provide. We analyzed the word-length distribution of these instructor-provided answers and used it to calibrate the model outputs. Specifically, the interquartile range spanned 50-108 words, with a mean of approximately

in Supplementary Information A.1. Before evaluation, all human and model answers were lightly standardized for typography and formatting to reduce superficial cues; tone markers commonly associated with LLMs (e.g., words of encouragement or praise for insightfulness) were retained, as they were often found in human answers as well.

In the last step, instructors in the role of judges completed individual evaluation sessions on a custom web interface inspired by the Chatbot Arena framework (23). Each trial displayed a Contracts question with two anonymized answers side-by-side. Pairs were either human-vs-LLM or LLM-vs-LLM; instructors never judged their own answers. Judges indicated which response they would prefer to give a student and could optionally (i) flag any answer–one or both in a pair–as "Low Quality: Harms student learning," and (ii) leave a short justification. To increase information per judgment, we used a forced choice setup with no-tie option; when two answers felt equivalent, instructors were asked to register their slight preference. Across the study, each instructor judged 150-200 comparisons yielding 2,918 total evaluations. Blinding and randomization governed source identity and left/right answer positioning.

Metrics

Our primary outcome of interest is expressed preference during the pairwise matchups. Secondary outcomes of interest focus on harmful responses, expressed either through the corresponding label or qualitative comments. We summarize performance using win rates for each model and instructor, computed over head-to-head comparisons with Wilson score intervals for uncertainty. We also computed Bradley-Terry rankings (36), a standard approach for converting pairwise preferences into overall rankings, with similar results reported in Supplementary Information D. Rates of harmful answers are aggregated in three steps: (1) if a judge flagged an answer as harmful at least once, all of that judge's evaluations of that answer were treated as harmful; (2) an answer's harmfulness rate was computed by averaging this quantity across all judges; and (3) a model/instructor's harmfulness rate is the average of their answers' harmfulness rates, with bootstrap resampling over answers for inference. We also report each judge's LLM-preference rate in human-vs-LLM pairwise comparisons. Complete descriptions of all statistical analyses, along with additional robustness checks, are available in Supplementary Information B.

We also assessed whether judges applied a shared professional standard by studying agreement on overlapping evaluations. The implicit null hypothesis is that, once we condition on judges' shared tendency to prefer LLM answers, their choices on any given pair are otherwise

90 words and an observed maximum of 155 words. These ranges informed the length constraints we applied when prompting the models.

uncoordinated—meaning residual agreement reflects only private taste rather than a shared evaluative criterion. To test this hypothesis, we focused on Gemini-vs-professor comparisons that were seen by at least two judges (we chose Gemini as the stronger-performing model, based on our Bradley-Terry assessment). For each such pair, computed the observed dyadic agreement rate. We then compared this rate to an arithmetic lower bound on agreement that is mechanically consistent with the marginal LLM win rate for the same pair. If judges agreed only as much as their shared LLM-preference necessitates, observed agreement should sit close to this bound; systematic excess of observed agreement over the bound is consistent with judges applying a common evaluative criterion beyond their shared LLM-preference. We then aggregated these quantities to the question level, using weights proportional to the number of unique judge dyads encountering a given pair of responses. Plotting observed agreement against this lower bound (Figure G.1) allows us to infer whether judges' quality standards converge beyond a mere preference for LLMs. This methodology targets convergence toward a generic, course-agnostic professional standard; it is important to note that it does not attempt to model or optimize for any one instructor's distinctive preferences. Full construction details and robustness checks are provided in Supplementary Information G.

To assess how the LLM preference interacts with the quality both of the human answer and the judge, we compared judges' choices in LLM-human matchups to two quantities: (i) a proxy for each judge's own quality as a writer—their average win rate when their answers were evaluated by others—and (ii) the quality of the human answer they were judging, measured as that answer's win rate computed in a leave-one-judge-out (LOJO) manner that excludes that judge's own decisions. We then examined how often the judge chose the human answer as a function of these two metrics (Figure H.2) and summarized judge-level calibration using within-judge Spearman correlations between LOJO human answer strength and LLM wins plotted against each judge's own win rate as an instructor (details can be found in Supplementary Information H, and Figure H.1). This design isolates whether better judges are more likely to favor stronger human answers without mechanically inflating the strength metric with their own vote.

To assess whether the LLM advantage can be explained by markers of style vis-a-vis content, we engineered lexico-syntactic features representing structural organization, reasoning nuance, legal anchors, confidence tone, clarity, pedagogical support, and length— guided by the AI tutoring literature (20; 27; 29; 37). We standardized features at the answer level, formed delta features (answer A minus answer B), and fit logistic models predicting which answer was preferred, using two-way clustered standard errors at the judge and answer pair level (details on Supplementary Information I, Figure I.1). To evaluate explanatory power, we derived predicted win probabilities from the fitted model and compared the observed

against the predicted win rates in deciles of predicted probability, separately by matchup (each LLM vs. instructor; Figure 3), with Wilson intervals for uncertainty.

LLM-as-judge: validation and deployment

Using an LLM as a stand-in for human evaluators is now a common methodology for scaling pairwise preference evaluation of open-ended generation (31–33). To evaluate *Llama-4 Maverick* as a judge, we used a standardized prompt that instructs the model to act as an experienced U.S. Contracts professor, weigh doctrinal accuracy, clarity, and pedagogical usefulness to approximate the instructions provided to professors (see Supplementary Information J for details).

The approach has documented limitations such as position and verbosity bias, as well as a tendency to favor outputs from models in the judge's own family (34; 35). We thus explicitly validate our chosen judge against human evaluators before deploying it. In particular, to assess *Maverick* aligned with human evaluators, we conducted a leave-one-out analysis across the 754 unique answer pairs seen by all 16 instructors. For each instructor, we removed their vote on every pair they judged, computed the majority decision among the remaining instructors, and compared both the instructor's own decision and the LLM-as-judge decision to this leave-one-out majority. Pairs with no remaining judges or with ties were excluded, typically removing about 14% of items per instructor. Because the LLM-as-a-judge and each instructor were evaluated on exactly the same set of pairs, this design provided a clean, pairwise comparison of human-to-LLM alignment. Across instructors, the LLM-as-a-judge matched the leave-one-out majority as often as the most aligned human judge and exceeded most others as shown in Figure J.1, with additional details provided in Supplementary Information J.

After validation, we used the selected LLM judge to adjudicate 42,652 cross-model comparisons generated by pairing four answer versions per question across our evaluated systems. We then fit a Bradley–Terry (BT) model (36) to the derived winner–loser pairs to estimate latent strengths on a common scale. Parameters are identified under the constraint that rankings sum to zero, and 95% confidence intervals result from the observed Fisher information as explained in Supplementary Information D and J.

A Design details

A.1 Prompts

System Prompt for NotebookLM

You are a Law School Contracts Professor. Your main knowledge source is the provided textbook chapters. Do not cite cases that are not in the source. Answer student questions as you would during office hours or after class: brief, direct, and grounded in the text. Avoid bullet points, restating the question, or using fillers. Respond naturally, not sounding like you conducted research. Keep answers between 50–108 words, ideally around 90. You may go up to 155 words only if nuance requires it.

System Prompt for Gemini 2.5 Pro

You are a Law School Contracts Professor. Answer student questions as you would during office hours or after class: brief and direct. Avoid bullet points, restating the question, or using fillers. Respond naturally, not sounding like you conducted research. Keep answers between 50–108 words, ideally around 90. You may go up to 155 words only if nuance requires it.

A.2 Question Table

Table A.1: Selected Contracts Law Questions

Question ID	Category	Instructor Labels	Question
1	Recall: Case or Code	Clear Answer, In Casebook	Could you summarize the facts in Ever-Tite Roofing Corp. v. Green?
2	Recall: Case or Code	Clear Answer	In Lefkowitz v. Great Minneapolis Surplus Store, why did the court consider the second but not the first advertisement an offer?

Continued on next page

Question ID	Category	Instructor Labels	Question
3	Recall: Case or Code	Clear Answer, In Casebook	When might a court award reliance damages instead of expectation damages?
4	Recall: Case or Code	Clear Answer, False Premise	In Russell v Texas, since the acceptance occurred by the offeree continuing to use the land, shouldn't that be considered an offer for an acceptance by performance (unilateral contract) rather than an acceptance by a return promise (bilateral contract)?
5	Recall: Case or Code	In Casebook	When you compare Kirksey v. Kirksey, Hamer v. Sidway, Langer v. Superior Steel Corp, and Pennsy Supply Inc.v. American Ash Recycling Corp, how did the consideration requirements evolve?
6	Recall: Case or Code	Clear Answer	Why was there substantial reliance in Hamer v. Sidway if it was not that difficult for the nephew to give up tobacco and alcohol?
7	Recall: Case or Code	Clear Answer	Which Restatement provision discusses promissory estoppel, and what does that provision require the plaintiff to prove to assert promissory estoppel?

Continued on next page

Question ID	Category	Instructor Labels	Question
8	Recall: Case or Code	Clear Answer, In Casebook	In disputes about alleged ambiguity of contractual language, does it matter whether there is a controlling meaning? If so, how?
9	Recall: Case or Code	Clear Answer, In Casebook	If no contract is formed under UCC 2-207(1), but then the parties behave like they have a deal, does the UCC say they have a contract or not?
10	Recall: Case or Code	Clear Answer, In Casebook	Can you explain UCC 2-207?
11	Recall: Doctrine	Clear Answer, In Casebook	If an offeree responds to the offer with a counteroffer, does the original offer expire?
12	Recall: Doctrine	Absence of Clear Rule/Precedent, In Casebook	Can you please explain the difference between a conditional gift promise (not legally binding) and bargained for consideration (legally binding)?
13	Recall: Doctrine	Clear Answer, In Casebook	I don't understand mutual assent. What do we mean by "objective intent to be bound"? What differentiates it from subjective intent?

Continued on next page

Question ID	Category	Instructor Labels	Question
14	Recall: Doctrine	Vague Standard, In Casebook	Can you help me understand when a promise made out of "moral obligation" is enforceable despite there being no consideration?
15	Recall: Doctrine	False Premise	Can you explain when additional terms knock each other out under UCC 2-207(2)?
16	Recall: Doctrine	In Casebook	If a contract is void under the statute of frauds, can a party still seek to enforce it by claiming promissory estoppel?
17	Recall: Doctrine	Clear Answer, In Casebook	When can parol evidence be used to show there was a mutual mistake about the meaning of a term in a fully integrated agreement?
18	Recall: Doctrine	Clear Answer, In Casebook	In what circumstances is an offer irrevocable?
19	Recall: Doctrine	Clear Answer, In Casebook	What is the difference between restitution, reliance and expectation damages?
20	Recall: Doctrine	Clear Answer, In Casebook	When can an agreement be enforced without consideration?

Continued on next page

Question ID	Category	Instructor Labels	Question
21	Hypotheticals	Clear Answer	Consider the following variation on the facts of Ever-Tite Roofing Corp. v. Green. After loading the trucks, while on the way to Owner's house, Contractor gets a call from Rich Client asking for immediate roof-repair work, in exchange for a hefty sum. Contractor diverts the trucks to Rich Client's house. Owner who did not hire a different contractor is suing for breach of contract. Will Owner prevail? Would your answer change if the offer form – Contractor's form that specified Owner is offeror – allowed only for acceptance by performance, rather than for acceptance by promise or performance (as in the actual case)?

Continued on next page

Question ID	Category	Instructor Labels	Question
22	Hypotheticals	Clear Answer, In Casebook	A offers to sell B a valuable autographed poster of Sabrina Carpenter for just $20, but specifies that B can only accept by having the words "Contracts Are Cool" displayed in 100-foot sky-writing letters over the center of their city. B telegrams back, "I accept your offer." Do A and B have a contract?

Continued on next page

Question ID	Category	Instructor Labels	Question
23	Hypotheticals	Clear Answer	Alan contracts with Formula Two, a car dealer in Toledo, Ohio to purchase and import a Fiasco sportscar for $35,000. Formula 2 locates the car in Italy and spends $25,000 to acquire the car and $5,000 in shipping and import fees. After Formula 2 notifies Alan that the car has arrived, Alan says he could not get a loan for the purchase and will no longer pay for the car. Formula 2 lists the car for sale on an enthusiast car auction website on which rare and imported cars are regularly sold. Formula 2 pays $500 to list the car on the site and sells the car for $30,000 at the end of the auction. Formula 2 then sues Alan for damages. Assume that Alan can introduce evidence that the buyer of the Fiasco at the auction immediately resold it to a car collector in Salem Oregon for $40,000. What is Formula 2 entitled to recover?

Continued on next page

Question ID	Category	Instructor Labels	Question
24	Hypotheticals	Vague Standard, In Casebook	Concerned that Nephew has too little time to study in college because he is also working a part-time job, Aunt promises to pay Nephew a monthly stipend equal to his earnings from the part-time job if he quits the job. If Nephew quits his part-time job. Do Aunt and Nephew have a contract that is supported by consideration?
25	Hypotheticals	Clear Answer, In Casebook	Under the Uniform Commercial Code, if two merchants exchange term sheets that contain different terms, with each sheet also containing a clause stating that acceptance is expressly conditional on the other party's assent to their terms, has a contract formed?
26	Hypotheticals	Clear Answer, In Casebook	I promise $10,000 to the winner of a swim race to Alcatraz Island. The swimmers dive off the dock and are going strong toward the island. When they're about half way, I stand up on Fisherman's Wharf with my bullhorn and yell "I REVOKE!" Can the winner of the race insist on the prize?

Continued on next page

Question ID	Category	Instructor Labels	Question
27	Hypotheticals	Clear Answer, In Casebook	Imagine a scenario similar to Lucy v. Zehmer, except that when the Zehmers were whispering about the whole thing being a joke, Lucy overheard the whispers. Then, Lucy goes to his buddy and says, "They think they're playing a joke on me, but they haven't told me they're joking! I know contract law. And I'm entitled to treat their statements as a reasonable person would hear them. Until they tell me they're joking, I can get them into a contract on the Ferguson Farm alright." Would that change the outcome of the case?
28	Hypotheticals	Vague Standard	If the company in Ever-Tite had not hired workers and loaded the trucks with materials but instead made a list of materials that they would take to the roofing supplier would the court still have viewed its action as acceptance?

Continued on next page

Question ID	Category	Instructor Labels	Question
29	Hypotheticals	Clear Answer, In Casebook	You contract with your stock broker to buy a share of Apple stock for $500. When the price drops to $400 the following month, can you get out of the deal by saying, "I made a mistake; I thought the price was going to go up?"
30	Hypotheticals	Clear Answer, In Casebook	Biff agrees to sell a car to Allee for $1,000. Allee wires Biff $1,000. Meanwhile, Miguel gets wind of the sale and offers Biff $1,500 for the car. Biff sells the car to Miguel. Allee sues for breach of contract. Suppose the judge finds for Allee and awards restitution damages. What amount would Biff have to pay Allee? What if the judge awards disgorgement?
31	Policy	—	Why are punitive damages disfavored in contract law?
32	Policy	Vague Standard, In Casebook	Is there a principled rationale for relaxing the requirement of consideration for contract modifications?

Continued on next page

Question ID	Category	Instructor Labels	Question
33	Policy	Vague Standard, In Casebook	Why do we have the Statute of Frauds, when it seems to let some parties break an agreement that the other party may have relied on?
34	Policy	Absence of Clear Rule/Precedent	In Sullivan v. O'Connor, the court says: "It has been suggested on occasion that agreements between patients and physicians by which the physician undertakes to effect a cure or to bring about a given result should be declared unenforceable on grounds of public policy." What's the public policy argument?
35	Policy	In Casebook	How does Judge Posner understand the doctrine of good faith (see Market Street Associates v. Frey)?
36	Policy	In Casebook	Why is quantity but not price an essential term under the UCC?
37	Policy	Vague Standard	What is the justification for the U.C.C. 2-207 approach of attempting to find formation even when an offer and acceptance contain different terms?

Continued on next page

Question ID	Category	Instructor Labels	Question
38	Policy	Clear Answer, In Casebook	What are the policy considerations behind the promissory estoppel leading to a quasi contract?
39	Policy	Vague Standard, In Casebook	An offeror is said to be "master of the offer," yet an offeror cannot stipulate that silence shall count as acceptance. Why not?
40	Policy	Absence of Clear Rule/Precedent, In Casebook	Why let parties specify their damages in a liquidated damages clause? It seems like a natural province of the court to decide what it ought to order to resolve a dispute.

A.3 Instructions to Instructors

A.3.1 Step 1 Instruction

Question-Answer Submission

As a first step in this study, we would like to ask each of you to submit eight questions with model answers to us under this Google Form. We will draw on a subset of the questions you submit for our evaluation exercise. The questions are divided across areas, and we ask that you provide two questions for each of the four categories:

- *Recall — Case or Code:* The question primarily asks for a summary of facts or law stated in a particular case or in the U.C.C. or Restatement.
- *Recall — Doctrine:* The question asks for an explanation of one or more doctrinal concepts.
- *Hypos:* The question requires analysis of how the doctrine would apply to a novel set of facts rather than a mere restatement of the relevant rules.
- *Policy:* The question asks about policy rationales or implications of a legal rule.

A full taxonomy of our questions, including examples, can be found in the attached document. Please note that the taxonomy you will see has been slightly updated in response to our statistical power analysis.

As a reminder, we would like the questions to be reflective of the types of inquiries that students commonly make after class or during office hours. Please also try to vary the subject, complexity, and difficulty of the questions.

In addition to the question themselves, we would like you to submit short model answers for each of your questions. The model answers should be representative of how you would answer the question if it was asked in office hours or after class. Please do not conduct research or spend significant time thinking about your answer. Please be concise. Perhaps it is helpful to envision that you have a long line of students burning to ask their question after class and that you have an important faculty meeting to attend, giving you only limited time for each individual answer.

Last but not least, for each of your questions, we would like you to assign all labels that appropriately characterize the question / answer. Labels are not mutually exclusive. The options are:

- *Clear:* There is one clear answer or narrow set of non-conflicting possible answers to a question.
- *Unclear:* Standard: The answer is unclear because the question raises an issue which is not obviously resolved by the relevant legal standard.
- *Unclear:* No Clear Rule or Precedent: The answer is unclear because there is no or conflicting authority on an issue necessary to resolve the question.
- *Answer in Casebook:* The casebook (Ayres, Klass & Stone's Studies in Contract Law) provides sufficient basis for answering the question, though inference may be required and the answer may not be complete.
- *False Premise:* The question reflects a clear misunderstanding of some aspect of contract law which needs to be corrected.

A.3.2 Step 2 Instruction

During Step 1 of this study, many of you have submitted questions that represent the types of inquiries students typically make after class or during office hours. We have selected a subset of 40 of these questions for inclusion in our study. In selecting the questions, we

aimed to represent various levels of complexity, and sought to avoid niche subjects that may not be covered in every contracts class.

In the current Step 2 of the study, each participant answers an even smaller subset of 10 questions submitted by other faculty. We do not believe this is the case, but if we accidentally assigned a question that you submitted yourself during Step 1, please contact us immediately. Looking ahead, in Step 3, we will then blindly compare and score these answers against answers submitted by large language models.

Instructions for writing an answer (also included in the survey):

- The answers should be representative of how you would answer the question if it was asked in office hours or after class.
- Please be concise. We expect that each answer takes no more than 3 minutes to write down.
- It is ok for an answer to reflect uncertainty.
- Please do your best to avoid conducting any research or to spend significant time thinking about your answer. That said, we understand this is not always possible. For instance, the question may be about a case that you don't typically cover in class, and thus will need to look up in order to answer the question. For these types of scenarios, we have included two additional, self-explanatory labels that you can check if appropriate, in addition to providing your answer. These labels are:
  - "I don't cover the subject of this question in my class"
  - "To answer this question, I had to do research (e.g. look up case facts or relevant doctrine)"

A.3.3 Step 3 Instruction

During Step 1 of this study, we collected common contract law questions from you. During Step 2, you have answered these questions. Simultaneously, we also generated responses to these questions via LLMs. In Step 3, we will examine whether we as a contract law teacher collective prefer answers generated by contract law teachers over answers generated by LLMs.

We have created a website that will show you a contract law question and two possible answers. You will not know which answer (if any) was generated by an LLM and which answer (if any) was written by one of your colleagues. We ask that you state which answer you prefer. By that, we mean: Which response would you rather give to a student who

asks the question in office hours or after your 1L contracts class? Consider clarity, doctrinal accuracy, and pedagogical usefulness—but ultimately go with your holistic judgment.

You also have the option to indicate if any answer is of such poor quality that it would—in your view—hinder student learning. You can select this option for any answer you see, including both answers in a pair. We also give you the option to leave a comment explaining your decision, although you are free to leave that box empty.

To ease the process as much as possible, we have organized the pairs we are showing to you by the relevant question. This means that the first batch of answer pairs you see all relate to the same question. You will then see another batch of paired answers, all belonging to a new question. And so on.

Below, we list a few points that are important to keep in mind:

- The goal is for everyone to document 150 preferences. During testing, we found that, after an initial period of familiarization, documenting a preference took an average of 45 seconds. We thus hope you will be able to complete 150 answer pairs over the course of two hours or so.

- Each vote matters. In case you are willing to document more than 150 preferences, we would certainly greatly appreciate it, as this will yield greater statistical reliability to the results. For that purpose, each survey includes 50 additional bonus answer-pairs. Again, we don't expect you to indicate your preference for these, but would certainly appreciate it.

- It can be tempting to think that there are cues which clearly reveal the use of an LLM. For instance, LLMs have been trained to be encouraging, and one might think that words of encouragement or affirmation at the beginning of an answer (e.g. "This is a great question!") indicate an LLM response. But this is not the case. Perhaps motivated by our prompt to envision you are answering questions in office hours, many human instructors adopted similar practices in the written answers they provided to us. Indeed, because encouragement and similar cues were so commonplace, we opted to make no changes in that regard to LLM or instructor responses.

- There is no option to choose a tie. This can sometimes be frustrating (I certainly felt so during our trial runs). But please know that this is by design. Setting the rankings up in that way allows us to get more information from every single vote, which in turn means we can lower the number of votes we require from each of you. If two answers truly feel equivalent in quality, please make your best judgment and go with your slight preference–even if it feels like a coin toss. You can rest assured that, when this happens

across voters, it will show up in the aggregate data as near-random, signaling that the two options are effectively indistinguishable in quality.

- The position of each answer (A-left/B-right) is randomized and balanced across the study.
- To stay within the allotted 120 minutes, we encourage you to make use of the comment box only sparingly.

B Computational details of results

B.1 Win Rate Computation

Win rates are computed from head-to-head comparisons where either (i) a professor's answer was evaluated against an LLM's answer, or (ii) two LLMs were compared. Each instructor or model contributed multiple answers, and every answer could appear in several comparisons. Judges were required to select a winner in each case (no ties). Formally, for a given model or instructor m:

$$\text{Win Rate}_m = \frac{\sum_{r \in R_m} W_r}{\sum_{r \in R_m} C_r} \tag{1}$$

where R_m is the set of all responses authored by m, W_r is the number of wins for response r, and C_r is the total number of comparisons involving r.

Confidence intervals were estimated using the Wilson score interval:

$$\hat{p} = \frac{w}{n}, \qquad \text{CI}_{\text{Wilson}} = \frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \tag{2}$$

where w is the number of wins, n is the total comparisons, and z is the critical value from the standard normal distribution for 95% confidence.

B.2 Judge Preference Computation

To quantify individual instructors' preferences as a judge, we computed a judge-specific preference rate. For each instructor j, we considered only LLM vs Human matchups. Let W_j denote the number of times instructor j selected the LLM response, and C_j the total number of such comparisons they judged. The judge rate is defined as:

$$\text{Judge Rate}_j = \frac{W_j}{C_j}. \tag{3}$$

Wilson score intervals at the 95% confidence level were computed as described in Eq. 2.

C Excluding Answers to Unfamiliar Questions

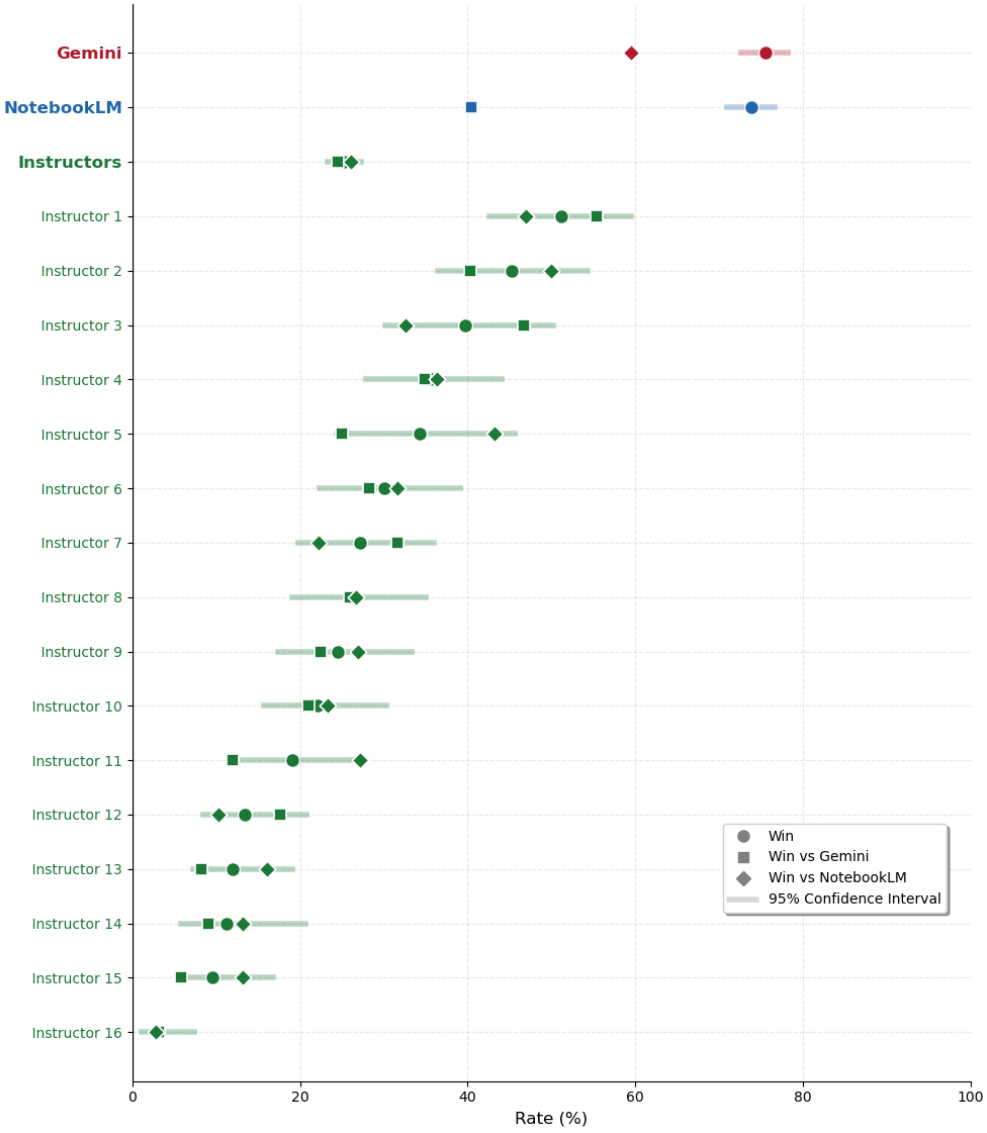


Figure C.1: Win rates after excluding any pair that contains an instructor answer flagged as either "I do not cover the subject of this question in my class" or "To answer this question, I had to do research (e.g., look up case facts or relevant doctrine)." Markers and 95% Wilson score intervals follow the same conventions as Figure 2

When instructors drafted answers to randomly selected questions, we asked them to be concise (targeting ≤ 3 minutes per answer) and to avoid conducting outside research or spending substantial time deliberating. We acknowledge that this would not always be feasible—for example, if a question concerned a case not covered in their course—and provided two additional, self-explanatory labels they could check alongside their answer when appropriate:

> "I do not cover the subject of this question in my class"
>
> "To answer this question, I had to do research (e.g., look up case facts or relevant doctrine)"

We recomputed win rates as described in Supplementary Information B, now excluding all comparisons that include at least one instructor answer flagging either label. This isolates potential uncertainty arising from questions outside an instructor's syllabus coverage/expertise or requiring ad hoc research. Only a small share of responses were flagged (14 of 153 unique instructor answers), so the exclusion has marginal effect: the resulting pattern in Figure C.1 is virtually unchanged from the overall results in Figure 2.

D Bradley-Terry Model

Our outcome is a set of pairwise preferences ("A" vs "B") collected under an incomplete and unbalanced design (not all respondents face every other on every question, and some pairs appear more often). Simple win rates combine strength with schedule imbalance and use incomparable denominators. The Bradley-Terry (BT) model (36) addresses this potential shortcoming by assigning each human/llm instructor m a latent "strength" ξ_m, and by then modeling head-to-head win probability as

$$\Pr(m \text{ beats } m') \;=\; \frac{1}{1 + e^{\xi_{m'} - \xi_m}}\,.$$

BT and its extensions are commonly employed for pairwise data (e.g., sports ratings, consumer choice, online A/B tests) because they recover a global scale even with sparse schedules. We follow this practice, using an implementation inspired by Chiang et al. (23).

Estimator. We collect all (winner, loser) pairs and maximize the BT log-likelihood. For identifiability, one coefficient is fixed to zero and we report centered estimates (sum-to-zero) for readability.

Uncertainty. Following Chiang et al. (23), we quantify uncertainty using the Huber-White's "sandwich" covariance for Bradley-Terry maximum likelihood estimations, yielding robust standard errors under mild conditions. From this covariance, we report two types of intervals: (i) ordinary per-parameter Wald intervals, and (ii) multiplicity-corrected intervals obtained by projecting χ^2 confidence ellipsoid for $\boldsymbol{\xi}$ onto each coordinate. The latter provide simultaneous 95% coverage for all coefficients and are the appropriate intervals for ranking claims.

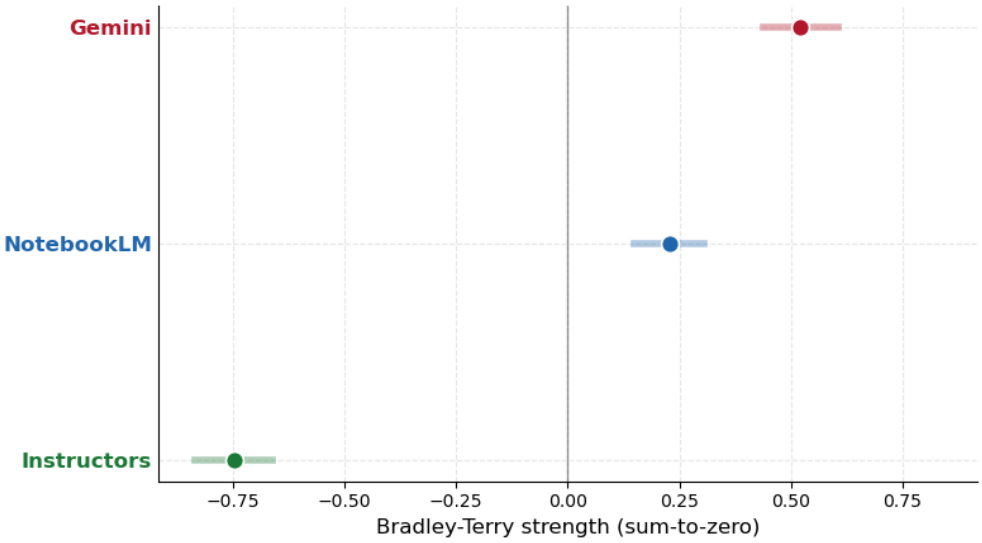
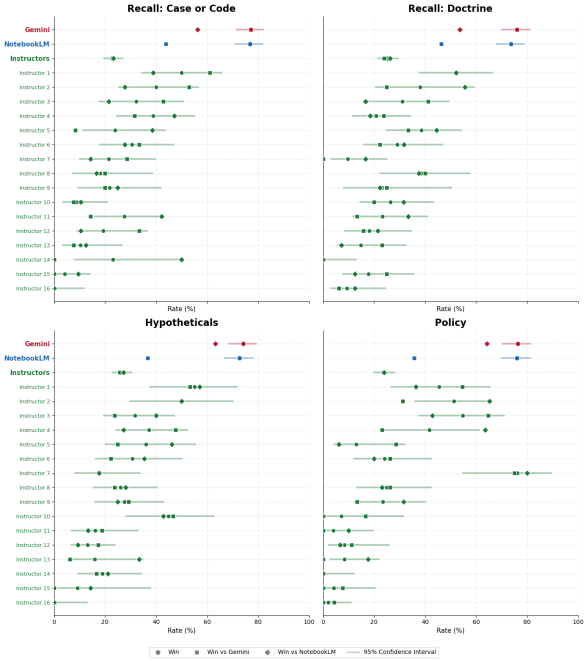


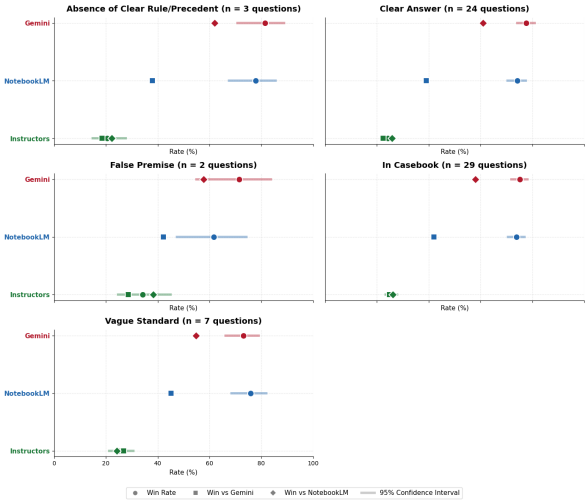
Figure D.1: Bradley–Terry (BT) log-odds strength estimates for Gemini, NotebookLM, and the instructor average. Points represent centered BT coefficients, interpreted as latent performance strengths in pairwise comparisons. Error bars show simultaneous multiplicity-corrected 95% intervals derived from projecting the χ^2 confidence ellipsoid for the full coefficient vector. Higher values indicate stronger estimated performance under the BT model.

Display and interpretation. We present the aggregated three-way comparison among Gemini, NotebookLM, and the instructor average. Figure D.1 reports BT point estimates (centered log-odds strengths) with simultaneous 95% confidence intervals.

E Win Rates per Category- and Subgroup-Level



(a) Category-level win rates with 95% Wilson intervals. Points: overall win rate; squares/-diamonds: head-to-head win rates vs Gemini and NotebookLM, respectively.



(b) Subgroup-level win rates for aggregated respondents (Gemini, NotebookLM, and Instructors) with 95% Wilson score intervals. Points mark the overall win rate within each subgroup. Squares and diamonds show head-to-head win rates versus Gemini and NotebookLM.

E.1 Win Rates per Category

Figure E.1a displays win rates by question category for Gemini (red), NotebookLM (blue), the aggregate of all instructors (green, bold), and each individual instructor (green, anonymized as *Instructor 1-16* in a fixed order based on the win rate performance from Figure 2. The four subplots correspond to the taxonomy used in the study: Recall-Case or Code, Recall-Doctrine, Hypotheticals, and Policy.

For every row, the point represents the overall win rate against all opponents within that category. If the row corresponds to an LLM, it represents the overall win rate against all instructors; if it corresponds to an instructor, it is computed using all matchups against LLMs. Horizontal bars denote 95% Wilson score intervals. Squares show head-to-head win rates vs Gemini and diamonds indicate head-to-head win rates vs NotebookLM for the same row. Win rates and confidence intervals were computed as shown in Supplementary Information B.

E.2 Win Rates by Question Subgroup

During question creation, instructors assigned one or more (non-mutually exclusive) subgroup labels to each question to characterize the question/answer. These subgroups were:

- *Clear Answer*: There is one clear answer or narrow set of non-conflicting possible answers to a question,
- *Vague Standard*: There is no clear answer because the question raises an issue which is not obviously resolved by the relevant legal standard,
- *Absence of Clear Rule/Precedent*: There is no clear answer because there is no or conflicting authority on an issue necessary to resolve the question,
- *In Casebook*: The casebook provides sufficient basis for answering the question, though inference may be required and the answer may not be complete,
- *False Premise*: The question reflects a clear misunderstanding of some aspect of contract law which needs to be corrected.

In total, 24 questions were labeled as Clear Answer, 2 as False Premise, 29 as In Casebook, 3 as Absence of Clear Rule/Precedent, and 7 as Vague Standard (counts do not sum to 40 because labels were not mutually exclusive).

Figure E.1b reports subgroup-specific win rates for Gemini (red), NotebookLM (blue), and the aggregate of all instructors (green). For each row, the point denotes the overall win

rate against all opponents within that subgroup (for LLM rows, against all instructors; for the Instructors row, against all LLMs). Horizontal bars indicate 95% Wilson score intervals. Squares show head-to-head win rates versus Gemini, and diamonds indicate head-to-head win rates vs NotebookLM. Win rates and confidence intervals were computed as shown in Supplementary Information B.

F Harmfulness Rate

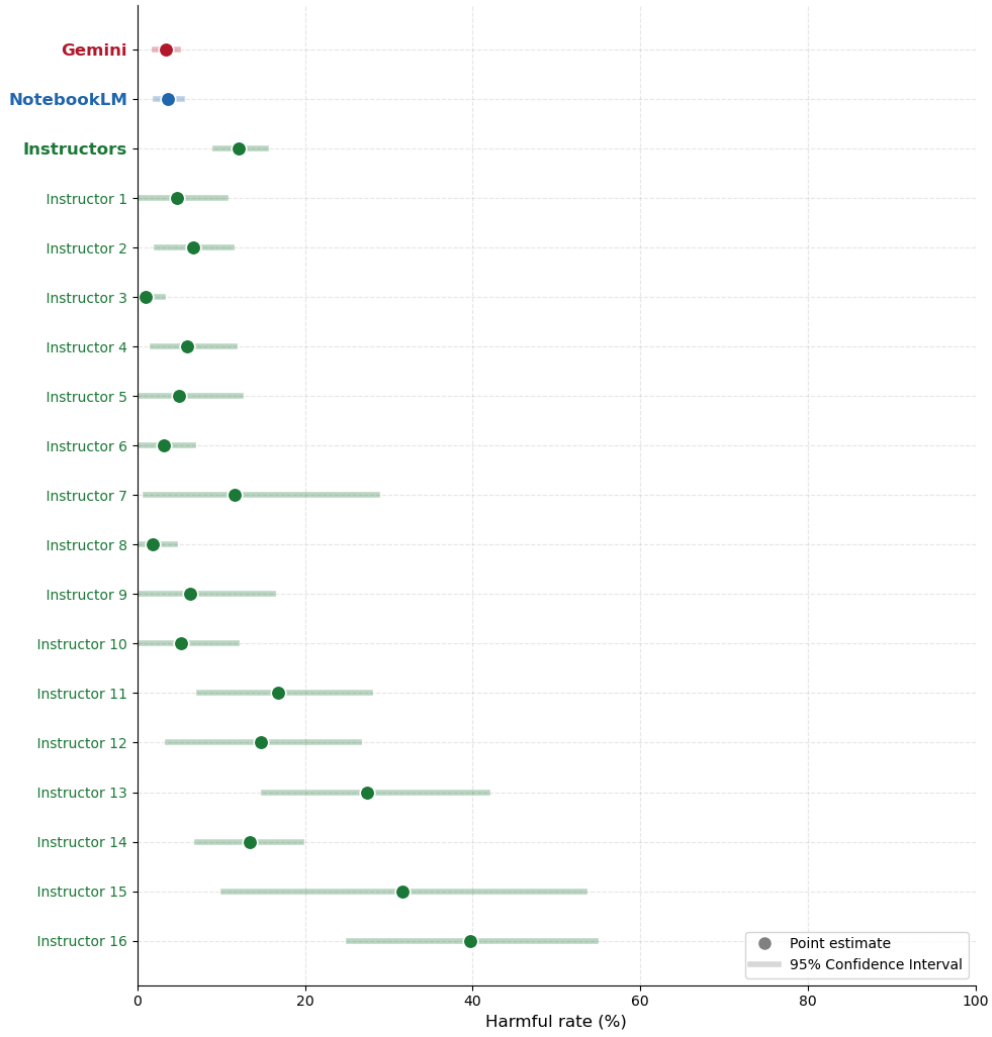


Figure F.1: Harmfulness rates for instructors and LLMs. Points denote mean harmfulness rates as defined in Supplementary Information F, with non-parametric 95% bootstrap confidence intervals obtained by resampling responses.

Harmfulness rates were computed in three steps:

1. **Judge-level marking**: if a judge labeled a response r as harmful at least once, then all instances of that response for that judge were considered harmful. This is done under the assumption that judges may find it superfluous to repeatedly label answers as harmful.

2. **Response-level aggregation**: for each response r, the harmfulness proportion was computed as

$$H_r = \frac{\text{Number of judges marking } r \text{ as harmful}}{\text{Total number of judges who evaluated } r} \tag{4}$$

3. **Model-level averaging**: the harmfulness rate for model/instructor m represents the mean across all responses:

$$\text{Harmful Rate}_m = \frac{1}{|R_m|} \sum_{r \in R_m} H_r \tag{5}$$

where R_m is again the set of responses authored by m.

Uncertainty in harmfulness rates was quantified using bootstrap resampling over responses, producing non-parametric confidence intervals.

G Intercoder Minimum

We quantify whether judges apply a shared standard by comparing observed dyadic agreement to the minimum agreement mechanically possibly by the LLM win rate. We restrict answer-pairs to those including responses from Gemini (as the stronger model) and instructors. For each answer pair j, let K_j be the number of judges who evaluated it and c_j the number who preferred the LLM; define $p_j = c_j/K_j$. Observed agreement for pair j is the fraction of agreeing judge dyads,

$$A_j \;=\; 1 - \frac{2\,c_j\,(K_j - c_j)}{K_j\,(K_j - 1)}\,,$$

and the pair-specific lower bound consistent with the marginal win rate is

$$L_j \;=\; \bigl|1 - 2p_j\bigr|\,,$$

which is the smallest possible dyadic agreement if judges merely follow idiosyncratic LLM-vs-human leanings with success probability p_j. We aggregate to the question level by weighting

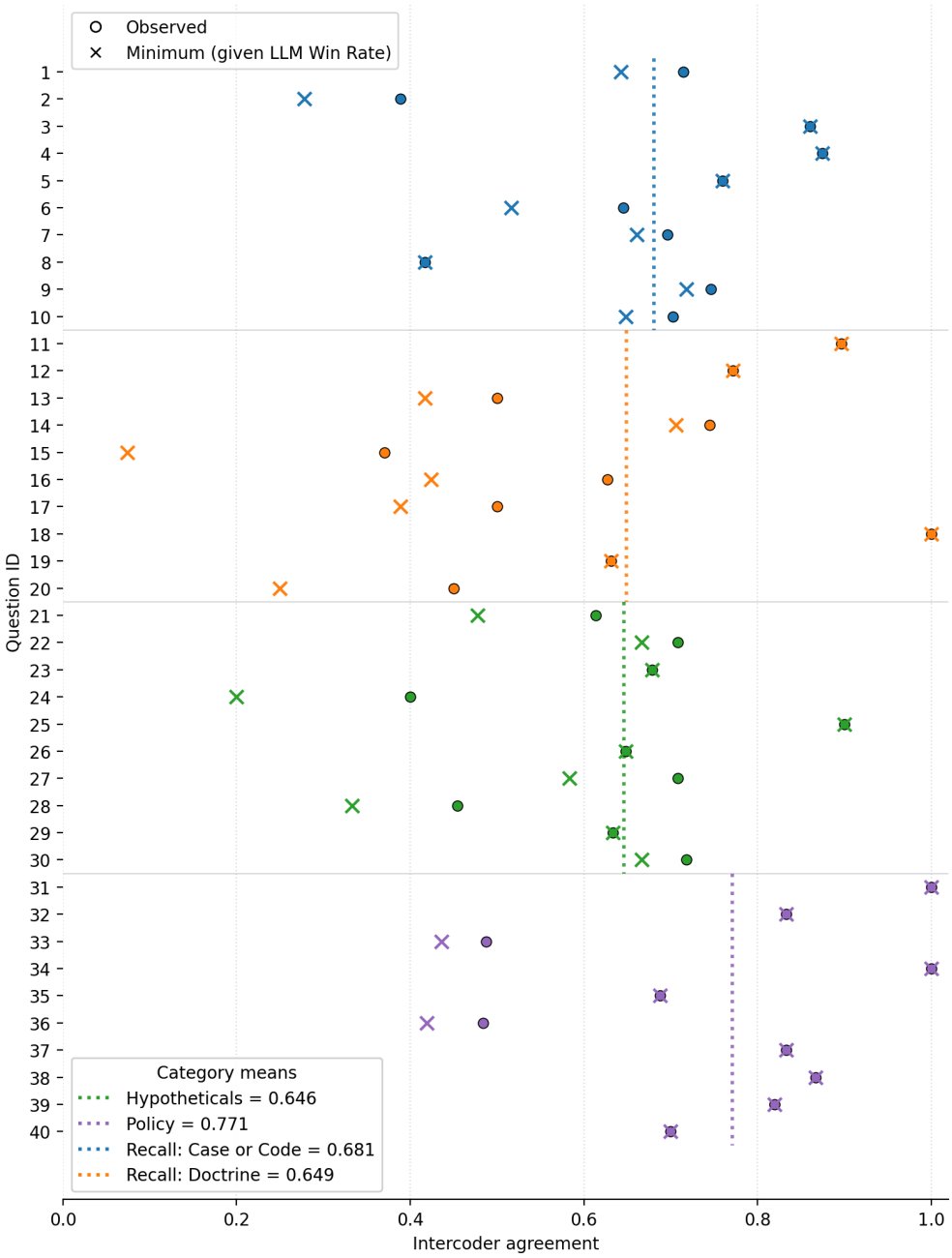


Figure G.1: Observed intercoder agreement by question with the minimum agreement implied by the LLM win rate. Each point shows the observed agreement rate for a question, computed on the subset of Gemini-vs-Instructor dyads, with colors indicating question category. Crosses indicate the smallest agreement rate considering the arithmetic constraints due to the observed LLM win rate.

each pair by its number of judge dyads $w_j = K_j(K_j - 1)/2$, yielding the dyad-weighted observed agreement $\bar{A}_q = \sum_j w_j A_j / \sum_j w_j$ and the corresponding dyad-consistent minimum $\bar{L}_q = \sum_j w_j L_j / \sum_j w_j$. Plotting $\bar{A}_q$ against $\bar{L}_q$ (Figure G.1) allows us to assess convergence beyond what is arithmetically implied by LLM win rates alone. Systematic excess of $\bar{A}_q$ over $\bar{L}_q$ indicates agreement consistent with a shared professional standard rather than purely private preferences.

H Instructor Strength and LLM Preference

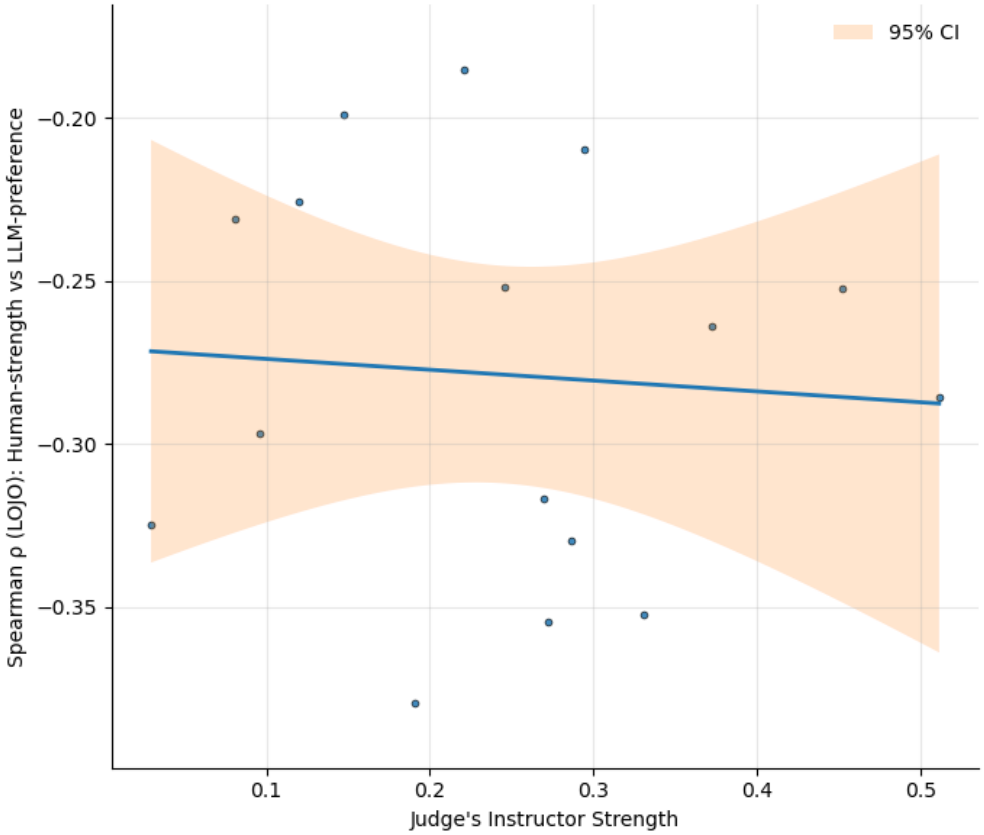


Figure H.1: Cross-judge trend between judges' own strength as an instructor and calibration to human answer quality in LLM vs human matchups. Each point represents a judge. The horizontal axis depicts the judge's global win rate when their own answers were evaluated (instructor strength). The vertical axis represents the within-judge Spearman correlation ρ between the leave-one-judge-out (LOJO) human-writer strength in each LLM vs human pair the judge rated and the LLM preference (an indicator that the LLM won, 1=LLM, 0=human). Negative ρ means the judge prefers human instructor responses when the judge is a stronger instructor. Line shows an OLS fit; shaded band is the 95% standard-error band for the mean trend.

Are 'strong' judges better at recognizing 'strong' human answers? Figure H.1 addresses whether "strong" judges, defined as those who themselves write answers with higher win probabilities, are better at recognizing strong, human-author instructor answers when deciding between a instructor and an LLM. The negative correlation values indicate that judges do tend to select stronger instructors, aligning with those whom other judges also deem strong. On the vertical axis, instructor strengths are computed using a leave-one-judge-out (LOJO) approach to prevent a judge's own ratings from inflating the metric of strength used to evaluate that same judge. Overall, there is no strong correlation between instructor quality and ability to identify and prefer other strong instructors.

H.1 Judge's own instructor win rate

For the horizontal axis, we measure each judge j's strength as an instructor using the same win-rate definition as Eq. 1, but restricted to responses authored by j:

$$X_j \;\equiv\; \mathrm{WinRate}_j^{\mathrm{writer}} \;=\; \frac{\sum_{r\in R_j} W_r}{\sum_{r\in R_j} C_r}, \tag{6}$$

where R_j is the set of responses authored by judge j, W_r is the number of wins for response r, and C_r is the number of comparisons involving r.

H.2 Spearman ρ LOJO

For the vertical axis, we computed the LOJO instructor strength with the following considerations in mind. As previously noted, we wanted to avoid circularity when analyzing judge j. Thus, we score the instructor h in a pair using a LOJO win rate that excludes all decisions made by judge j:

$$\theta_h^{(-j)} \;=\; \frac{\sum_{r\in R_h}\big(W_r - W_{rj}\big)}{\sum_{r\in R_h}\big(C_r - C_{rj}\big)}\,, \tag{7}$$

where R_h is the set of responses authored by human instructor h, and W_{rj} and C_{rj} are, respectively, the number of wins and comparisons for response r specifically in decisions by judge j.

Furthermore, let I_j be the set of LLM vs. instructor comparisons judged by j. For each $i \in \mathcal{I}_j$, let $h(i)$ denote the human instructor in that pair and define a binary indicator of whether the LLM won:

$$L_{ji} \;=\; \mathbb{1}\{\text{LLM wins in comparison } i \text{ judged by } j\}\,. \tag{8}$$

We then compute the within-judge Spearman correlation between the LOJO instructor strength and LLM wins:

$$Y_k \;\equiv\; \rho_j \;=\; \text{Spearman}\Big(\{\theta^{(-j)}_{h(i)} : i \in \mathcal{I}_j\}, \{L_{ji} : i \in \mathcal{I}_j\}\Big). \tag{9}$$

As previously stated, a $\rho < 0$ indicates that the judge is less likely to select the LLM when the instructor in a pair is stronger (as defined by the pool of judges), which we interpret as consistent/coherent behavior across judges towards a group definition of instructor quality.

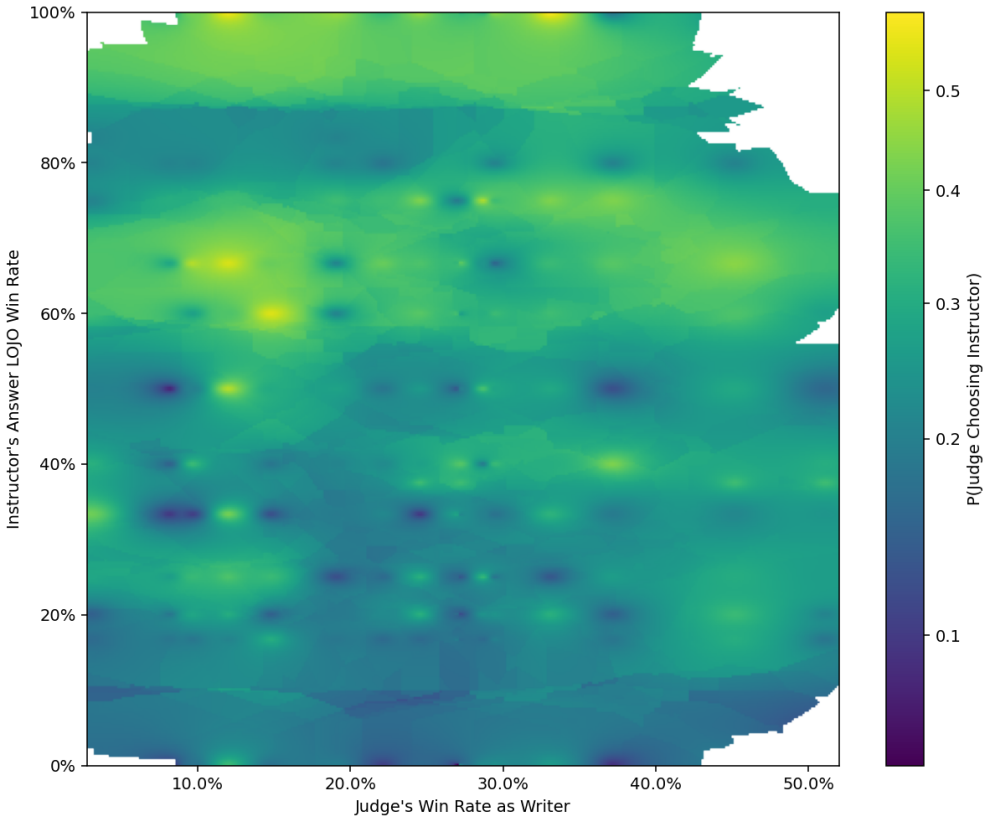


Figure H.2: Judge level sensitivity to human instructor vs. LLM answers. Heatmap shows the estimated probability that a judge chooses the instructor over an LLM as a function of (horizontal axis) the judge's own strength as a writer (win rate) and (vertical axis) the human answer's strength measured as a leave-one-judge-out (LOJO) win rate for that human answer within the same pair. Colors indicate the probability of preferring the human instructor answer; exact-match cells use empirical values, and the surface elsewhere uses Empirical-Bayes-smoothed, count-weighted inverse-distance interpolation.

Figure H.2 shows the above-detailed (B) instructor's win rate on the horizontal axis and the aforementioned LOJO instructor win rate on the y-axis. We estimated the surface by aggregating identical (x, y) coordinates and computed the observed share of times the

instructor was preferred. To stabilize sparse cells, we applied Empirical-Bayes shrinkage toward the global mean, then produced a continuous surface using count-weighted inverse distance weighting (IDW) on a grid, masking regions too far from observed data.

I Feature Importance Analysis

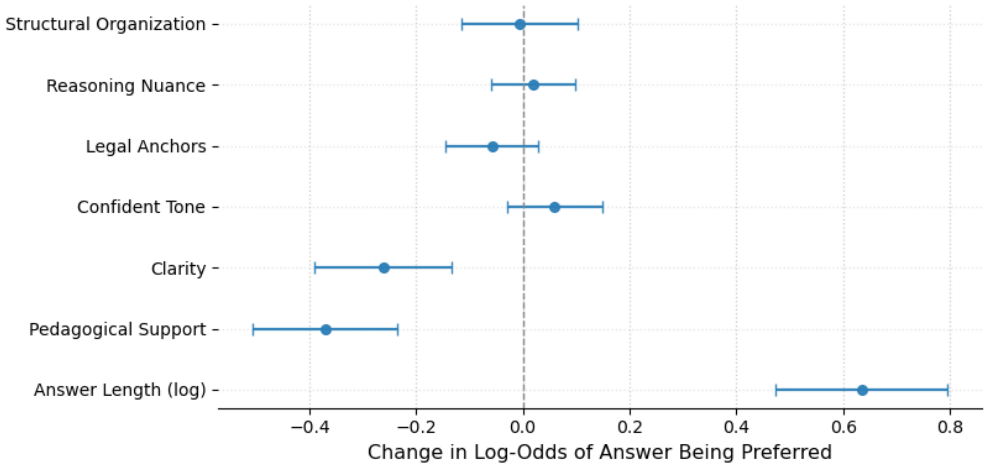


Figure I.1: Estimated coefficients from a logistic regression of pairwise answer choices on differences in textual and pedagogical features, with two-way clustered standard errors (by judge and answer pair). Dots represent point estimates; horizontal lines show 95% confidence intervals. Positive values indicate that higher feature values increase the likelihood of an answer being preferred. Answer length emerges as the strongest predictor.

We modeled judges' pairwise choices with a logistic regression using only Δ features—feature differences between answer A and answer B. Specifically, our features consist of legal anchors, reasoning nuance, clarity, confident tone, structural organization, pedagogical support, and log length. The binary outcome was $y = 1$ if A was preferred, 0 otherwise. We fit the logit without fixed effects, although our results are not sensitive to their addition. We then computed two-way cluster-robust standard errors by judge and by question-answer pair. Figure I.1 displays each Δ feature's coefficient and 95% Confidence Interval on the log-odds scale. Positive coefficients indicate that increasing a feature in A relative to B raises the odds that A is chosen; negative coefficients indicate the opposite. These estimates both explain which textual attributes move judge preferences and produce the predicted probabilities used for downstream evaluation (see below). Table I.1 shows the estimated coefficients, as well as indicating that features length, pedagogical support (the sum of the rate of question marks and the scaffolding rate per answer (39; 40)), and clarity (Flesch reading-ease score (38)) are statistically significant.

Table I.1: Logit coefficients (Two-way clustered SEs). Coefficients on log-odds scale for each Δ feature.

Term	Coef	SE	p	CI low	CI high	Sig
Length log	0.634	0.082	1.02×10^{-14}	0.474	0.795	***
Pedagogical Support	-0.371	0.069	7.68×10^{-8}	-0.506	-0.236	***
Clarity	-0.262	0.066	6.92×10^{-5}	-0.391	-0.133	***
Confidence Tone	0.060	0.045	0.186	-0.029	0.148	
Legal Anchors	-0.058	0.044	0.189	-0.145	0.029	
Reasoning Nuance	0.018	0.040	0.650	-0.061	0.097	
Structural Organization	-0.006	0.055	0.911	-0.115	0.102	

We restricted the data to LLM vs. instructor comparisons. For each row, we aligned predictions and outcomes to the focal LLM: if the model was on side A, we used its predicted win probability as p_A; if on side B, then $1 - p_A$. Within each matchup, we formed deciles and computed, per decile, the mean predicted win rate, the observed win rate, and a 95% Wilson confidence interval. Figure 3 plots on the observed win rate—with errors bars against the binned mean predicted win rate , overlaying the $y = x$ line representing perfect calibration, and a 50% reference line. Points near the diagonal indicate well-calibrated probabilities.

J LLM-as-a-Judge

J.1 Prompt

System Prompt for LLM-as-a-Judge

You are an experienced U.S. contract-law professor. Your task is to read a law-school QUESTION plus two ANSWERS ("A" and "B") and decide which answer you would prefer to give to a 1L student. Judge each answer on the following, weighting them roughly equally unless one dimension is clearly decisive:

1. Doctrinal accuracy: correctness and completeness of the legal analysis.
2. Clarity: logical flow, plain language, and concision.
3. Pedagogical usefulness: capacity to help a novice understand and apply the doctrine.

Additional instructions:

- You do not know who authored either answer. Ignore stylistic cues that might reveal an LLM or a human.
- No ties. If the answers feel identical, pick the slightly better one.
- Think through the criteria step-by-step privately. After you finish reasoning, output only the JSON object shown below.
- Do not reveal or quote your private reasoning.

JSON example format: "preferred":"A"—"B"
BEGIN SCRATCHPAD (private)

J.2 Choosing our LLM-as-a-judge: Leave-one-out agreement analysis between human instructors and the LLM-as-a-judge

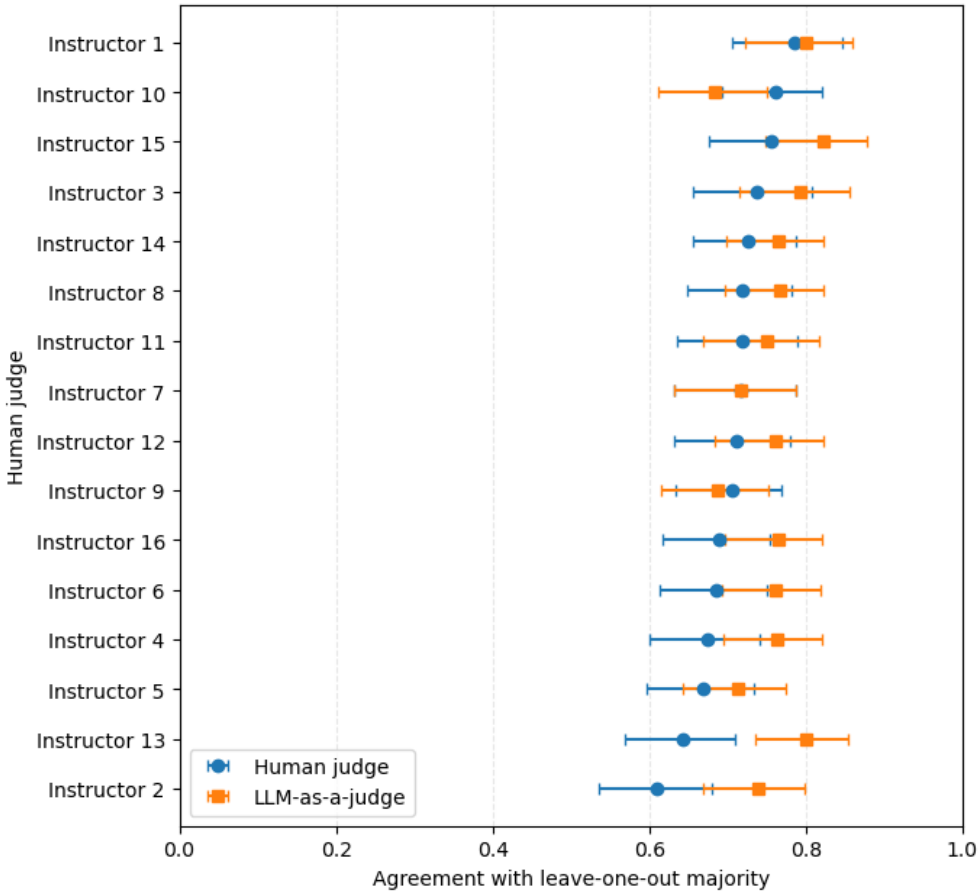


Figure J.1: Leave-one-out agreement with the human consensus for each instructor and the LLM-as-a-judge. Blue circles show the proportion of times that the instructor's decision agrees with this leave-one-out majority, and orange squares show the corresponding agreement rate for the LLM evaluated on the same set of pairs. Horizontal bars indicate 95% Wilson score confidence intervals for each proportion. Instructors are anonymized as "Instructor 1"–"Instructor 16" based on their overall win rate ranking in the main pairwise-comparison experiment, with 1 being best and 16 worst.

To compare how closely each human instructor and the LLM-as-a-judge align with the rest of the judges, we conducted a leave-one-out majority-vote analysis on the pairwise comparison data. Let items $i = 1, \ldots, N$ index the 754 unique answer pairs seen by the $J = 16$ instructors. Each human judge j provides a label $y_{ij} \in \{A, B\}$ the LLM judge produces $\hat{y}_i \in \{A, B\}$. Naturally, the LLM decision is constant within each unique pair, whereas human decisions vary across judges.

For each instructor j, we restricted attention to the set of unique answer pairs for which

the human provided a judgment. For every such pair, we first removed j's decision and computed a leave-one-out consensus among the remaining human judges. Concretely, we counted how many of the remaining judges chose said answer versus the remaining option; the one with the larger count was treated as the majority winner for that pair. Pairs for which (i) no other judge was available or (ii) there was an exact tie between the two options were excluded from the analysis for that instructor (anecdotally, the average unique pairs dropped rate was 13.78% across instructors and it was never higher than 17.20%.)

For the subset of pairs with a well-defined leave-one-out majority, we then computed two agreement rates:

1. the proportion of pairs for which instructor j's decision matches the leave-one-out majority; and

2. the proportion of pairs for which the LLM-as-a-judge's decision matches the same leave-one-out majority.

These two quantities are computed on exactly the same set of pairs for each instructor, so that the human and LLM are evaluated on an identical decision set.

For each proportion, we construct a 95% confidence interval using the Wilson score interval, as calculated in Supplementary Information B.

To preserve anonymity and to ensure consistency across figures, we mapped each instructor to their overall win rate ranking in the main pairwise-comparison experiment, with "Instructor 1" having the highest win rate and "Instructor 16" the lowest.

Figure J.1 shows that the selected LLM judge (*Llama-4 Maverick*) matches the alignment of the top-performing human judge and, overall, exceeds the alignment of most human judges with respect to the leave-one-out majority.

J.3 Scaling analysis with LLM-as-a-judge

We generated four answers per question for each model, created all cross-model pairings (totaling 42,652 unique pairs), and judged them with the validated LLM. We aggregated at the base-model level and computed BT strengths and their confidence intervals as shown in Supplementary Information D. Results can be found in Figure 4.

K Representativeness of the Instructor Sample

The sixteen instructors who participated in the study were drawn from a pool of sixty professors who used the casebook (54) to teach first-year Contracts in the four years preceding

the study. To assess how the respondent sample compares to the full pool, we collected publicly available characteristics on all sixty professors along four dimensions: gender, law school ranking (per the US News 2026 rankings[LINK]), tenure status, and years since tenure. Law school rankings were retrieved from the public US News listing; tenure status and years since tenure were determined from publicly available faculty biographies and CVs. Where a tenure year could not be confidently established from public sources, the professor is excluded from the years-since-tenure summary (but retained for all other dimensions); the relevant denominators are reported in Table K.1.

Table K.1: Comparison of respondents (n=16) to non-respondents (n=44) and the full casebook-adopter pool (n=60) along observable characteristics. Counts (and column percentages) shown for each dimension; the row reporting years since tenure shows the median among those with a confidently identifiable tenure year, with n noted in parentheses.

Dimension	Respondents (n=16)	Non-respondents (n=44)	Full pool (n=60)
Gender (% female)	4 (25%)	16 (36%)	20 (33%)
Law school rank (US News 2026)			
T14 (1–14)	6 (38%)	7 (16%)	13 (22%)
T15–50	2 (13%)	5 (11%)	7 (12%)
T51–100	3 (19%)	13 (30%)	16 (27%)
T101+	4 (25%)	17 (39%)	21 (35%)
Unranked / NA	1 (6%)	2 (5%)	3 (5%)
Tenured	15 (94%)	35 (80%)	50 (83%)
Years since tenure (median, of those with a known tenure year)	16 yrs (n=15)	20 yrs (n=35)	19 yrs (n=50)

Table K.1 reports the joint comparison. The respondent sample over-represents professors at T14 law schools (38% vs. 22% in the full pool) and, more modestly, under-represents female professors (25% vs. 33%). Respondents are also somewhat more likely to be tenured (94% vs. 83%), though their median time since tenure is shorter (16 vs. 19 years). We did not collect characteristics that would have required private outreach to the non-respondent group (e.g., teaching experience specific to Contracts, geographic background beyond institutional location), and we report the dimensions for which we could obtain consistent public data on

both respondents and non-respondents.

References

[1] Eva Verhelst, Ruben Janssens, Thomas Demeester, and Tony Belpaeme. Adaptive second language tutoring using generative ai and a social robot. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, page 1080–1084, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703232. doi: 10.1145/3610978.3640559. URL `https://doi.org/10.1145/3610978.3640559`.

[2] G. Kestin, K. Miller, A. Klales, et al. Ai tutoring outperforms in-class active learning: An rct introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15:17458, 2025. doi: 10.1038/s41598-025-97652-6. URL `https://doi.org/10.1038/s41598-025-97652-6`.

[3] Michael Vaccaro Jr, Mikayla Friday, and Arash Zaghi. Multi-agentic llms for personalizing stem texts. *Applied Sciences*, 15(13):7579, 2025.

[4] Byung Ok Kang, Hyung-Bae Jeon, and Yun Kyung Lee. Ai-based language tutoring systems with end-to-end automatic speech recognition and proficiency evaluation. *ETRI Journal*, 46(1):48–58, 2024.

[5] Ali M Fazlollahi, Mohamad Bakhaidar, Ahmad Alsayegh, Recai Yilmaz, Alexander Winkler-Schwartz, Nykan Mirchi, Ian Langleben, Nicole Ledwos, Abdulrahman J Sabbagh, Khalid Bajunaid, et al. Effect of artificial intelligence tutoring vs expert instruction on learning simulated surgical skills among medical students: a randomized clinical trial. *JAMA network open*, 5(2):e2149008–e2149008, 2022.

[6] Vivianna Fang He, Sihan Li, Phanish Puranam, and Feng Lin. Tool or tutor? experimental evidence from ai deployment in cancer diagnosis. *arXiv preprint arXiv:2502.16411*, 2025.

[7] Adit Gupta, Jennifer Reddig, Tommaso Calo, Daniel Weitekamp, and Christopher J MacLellan. Beyond final answers: Evaluating large language models for math tutoring. In *International Conference on Artificial Intelligence in Education*, pages 323–337. Springer, 2025.

[8] Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, et al. Llm-as-a-tutor in efl writing education: Focusing on evaluation of student-llm interaction. *arXiv preprint arXiv:2310.05191*, 2023.

[9] Ambroise Baillifard, Maxime Gabella, Pamela Banta Lavenex, and Corinna S Martarelli. Implementing learning principles with a personal ai tutor: A case study. *arXiv preprint arXiv:2309.13060*, 2023.

[10] Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of llm tutors. *arXiv preprint arXiv:2502.18940*, 2025.

[11] Kaushal Kumar Maurya, KV Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. Unifying ai tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of llm-powered ai tutors. *arXiv preprint arXiv:2412.09416*, 2024.

[12] Michael P Lynch. *Truth in context: An essay on pluralism and objectivity*. MIT press, 1998.

[13] Gerard E Lynch. Complexity, judgment, and restraint. *NYUL Rev.*, 95:621, 2020.

[14] Chloë J Wallace. The pedagogy of legal reasoning: democracy, discourse and community. *The Law Teacher*, 52(3):260–271, 2018.

[15] Wilson R Huhn. Teaching legal analysis using a pluralistic model of law. *Gonz. L. Rev.*, 36:433, 2000.

[16] Lisa Larrimore Ouellette, Amy Motomura, Jason Reinecke, and Jonathan S. Masur. Can ai hold office hours? *Journal of Legal Education (forthcoming)*, February 4 2025. doi: 10.2139/ssrn.5166938. URL `https://ssrn.com/abstract=5166938`. Stanford Public Law Working Paper; University of Chicago Law School, Coase-Sandor Institute for Law & Economics Research Paper No. 25-17; University of Chicago Law School, Public Law & Legal Theory Research Paper No. 25-13.

[17] Jonathan H. Choi, Kristin E. Hickman, Amy Monahan, and Daniel Schwarcz. Chatgpt goes to law school. *Journal of Legal Education*, 71:387, January 2023. doi: 10.2139/ssrn.4335905. URL `https://ssrn.com/abstract=4335905`. Available at SSRN: `https://ssrn.com/abstract=4335905`.

[18] Jonathan H. Choi and Daniel Schwarcz. Ai assistance in legal analysis: An empirical study. *Journal of Legal Education*, 73, August 2023. doi: 10.2139/ssrn.4539836. URL `https://ssrn.com/abstract=4539836`. Forthcoming; available at SSRN: `https://ssrn.com/abstract=4539836`.

[19] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382 (2270):20230254, 2024.

[20] LearnLM Team, Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Brett Wiltshire, Brian Veprek, Daniel Gillick, Daniel Kasenberg, Derek Ahmed, et al. Learnlm: Improving gemini for learning. *arXiv preprint arXiv:2412.16429*, 2024.

[21] John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, Zechariah Zhu, Jessica B. Kelley, Dennis J. Faix, Aaron M. Goodman, Christopher A. Longhurst, Michael Hogarth, and Davey M. Smith. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6):589–596, 06 2023. ISSN 2168-6106. doi: 10.1001/jamainternmed.2023.1838. URL `https://doi.org/10.1001/jamainternmed.2023.1838`.

[22] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[23] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

[24] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

[25] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[26] Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simon Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. Kim, Patrick Chao, Samuel Miserendino, Gildas Chabot, David Li, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. Gdpval: Evaluating ai model performance on real-world economically valuable tasks. *Preprint*, 2025. Not yet published. Available at `https://cdn.openai.com/pdf/d5eb7428-c4e9-4a33-bd86-86dd4bcf12ce/GDPval.pdf`.

[27] Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*, 2024.

[28] Hanyi Xu, Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Philip S Yu. Large language models for education: A survey. *arXiv preprint arXiv:2405.13001*, 2024.

[29] Hyein Seo, Taewook Hwang, Jeesu Jung, Hyeonseok Kang, Hyuk Namgoong, Yohan Lee, and Sangkeun Jung. Large language models as evaluators in education: Verification of feedback consistency and accuracy. *Applied Sciences (2076-3417)*, 15(2), 2025.

[30] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.

[31] Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069, 2023.

[32] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 2024.

[33] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36: 46595–46623, 2023.

[34] Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and

favor their own generations. *Advances in Neural Information Processing Systems*, 37: 68772–68802, 2024.

[35] Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. Justice or prejudice? quantifying biases in llm-as-a-judge. In *International Conference on Learning Representations*, volume 2025, pages 102351–102390, 2025.

[36] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[37] Irina Jurenka, Markus Kunesch, Kevin R McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, et al. Towards responsible development of generative ai for education: An evaluation-driven approach. *arXiv preprint arXiv:2407.12687*, 2024.

[38] Rudolf Flesch. How to write plain english. *University of Canterbury. Available at http://www. mang. canterbury. ac. nz/writing_guide/writing/flesch. shtml.[Retrieved 5 February 2016]*, 1979.

[39] Minoo Alemi. Functions and strategies of teachers' discursive scaffolding in english-medium content-based instruction. 2020.

[40] Janneke Van de Pol, Monique Volman, and Jos Beishuizen. Scaffolding in teacher–student interaction: A decade of research. *Educational psychology review*, 22(3):271–296, 2010.

[41] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.

[42] Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*, 2023.

[43] Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. The distracting effect: Understanding irrelevant passages in rag. *arXiv preprint arXiv:2505.06914*, 2025.

[44] Quinn Leng, Jacob Portes, Sam Havens, Matei Zaharia, and Michael Carbin. Long context rag performance of large language models. *arXiv preprint arXiv:2411.03538*, 2024.

[45] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16 (1):64–93, 2024.

[46] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242, 2025.

[47] Jonathan H Choi. Large language models are unreliable judges. *Available at SSRN 5188865*, 2025.

[48] Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*, 2025.

[49] Ivo Petrov, Jasper Dekoninck, Lyuben Baltadzhiev, Maria Drencheva, Kristian Minchev, Mislav Balunović, Nikola Jovanović, and Martin Vechev. Proof or bluff? evaluating llms on 2025 usa math olympiad. *arXiv preprint arXiv:2503.21934*, 2025.

[50] David Freeman Engstrom and Jonah B Gelbach. Legal tech, civil procedure, and the future of adversarialism. *University of Pennsylvania Law Review*, pages 1001–1099, 2021.

[51] Edward H Levi. *An introduction to legal reasoning*. University of Chicago Press, 2013.

[52] Shiri Melumad and Jin Ho Yun. Experimental evidence of the effects of large language models versus web search on depth of learning. *Available at SSRN 5104064*, 2025.

[53] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

[54] Ian Ayres, Gregory Klass, and Rebecca Stone. *Studies in Contract Law*. Foundation Press, St. Paul, MN, 10 edition, 2024. ISBN 9781647085445. CasebookPlus ISBN: 9798887867120.

[55] Ben Gomes. Learn in newer, deeper ways with gemini. Blog post on Google's "The Keyword", May 2025. URL `https://blog.google/outreach-initiatives/education/google-gemini-learnlm-update/`. Accessed: 2025-10-01.

[56] Raiza Martin and Steven Johnson. Introducing notebooklm. Blog post on Google's "The Keyword", Jul 2023. URL https://blog.google/technology/ai/notebooklm-google-ai. Accessed: 2025-10-01.