Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| REVIEW ARTICLE

Benchmarking Legal AI on Ambiguity: A New Evaluation Framework for Edge-Case Reasoning

Mitul Ashvinbhai Trivedi
The Walsh College, USA
Corresponding Author: Mitul Ashvinbhai Trivedi, **E-mail**: reachmitultrivedi@gmail.com

| ABSTRACT

Legal AI systems are increasingly deployed in contexts that demand nuanced interpretive judgment, yet existing evaluation benchmarks remain anchored to deterministic accuracy measures that reward single correct answers and penalize appropriate epistemic hedging. This misalignment between evaluation design and the realities of legal practice produces systems that perform well on benchmarks while failing in precisely the situations where professional legal judgment is most consequential.

This article proposes an Ambiguity-Aware Evaluation Framework for legal AI that moves beyond top-1 accuracy as the primary performance criterion. The framework is developed through a structured analysis of the structural limitations of current legal AI benchmarks, a taxonomy of legal question types, and the specification of novel evaluation metrics and architectural requirements suited to ambiguity-rich legal reasoning. The framework distinguishes three categories of legal questions: deterministic questions governed by clear statutory rules and uniform doctrine; contested questions characterized by legitimate interpretive disagreement across jurisdictions or scholarly opinion; and indeterminate questions requiring policy judgment, fact-sensitive reasoning, or analogical extension into unsettled legal territory. For each category, the framework specifies appropriate evaluation criteria and defines three novel metrics: ambiguity detection rate, which measures whether a system correctly identifies when a legal question lacks a settled answer; interpretation completeness, which evaluates whether a system enumerates the full space of legally viable positions; and qualification appropriateness, which assesses whether expressed confidence is calibrated to the underlying degree of legal certainty. The article further identifies the architectural components required to operationalize the framework, including retrieval-based ambiguity detection mechanisms, structured interpretation enumeration modules, and uncertainty calibration training objectives that reward hedged reasoning on contested and indeterminate questions rather than penalizing it. The findings indicate that current benchmark designs systematically incentivize overconfidence, fail to distinguish genuine legal reasoning from surface pattern matching, and inadequately prepare systems for deployment in professional legal contexts. The proposed framework provides a more practically valid standard for evaluating legal AI, one that reflects the epistemic demands of skilled legal practice and rewards systems capable of recognizing and communicating the limits of deterministic legal analysis.

| KEYWORDS

Ambiguity-Aware Evaluation, Legal AI Benchmarking, Interpretive Uncertainty, Contested Legal Questions, Epistemic Humility

| ARTICLE INFORMATION

ACCEPTED: 15 August 2026 **PUBLISHED:** 08 September 2026 **DOI:** 10.32996/jcsts.2026.8.9.2

1. Introduction

Legal AI benchmarks also lack diversity in the questions they ask, whereas in practice lawyers have to deal with ambiguous statutes and case law and fact patterns with multiple viable interpretations. In contrast, customary legal AI datasets mostly consist of single-answer tasks measured using top-1 accuracy. On these tasks, SOTA BERT-based models achieve over 48% top-1 accuracy. The SOTA model achieves 51% accuracy on specialized legal tasks, while the average human achieves 49% accuracy (Koenig et al., 2023). In practice, top-1 accuracy does not measure how well a legal AI would perform. For instance, legal uncertainty or doctrinal disagreement abounds in many cases in the appellate courts (Savelka et al., 2023). To address this gap, this article proposes an

Ambiguity-Aware Evaluation Framework to measure how well legal AI performs under legal uncertainty, interpretive disagreement, and hedged legal conclusions. The framework is presented to model a collection of legal question types, ranging from deterministic to contested to indeterminate legal questions, by novel measures, including the ambiguity detection rate, the thorough identification of all possible alternative interpretations, and the calibration of expressed uncertainty. The intention is to improve the evaluation of legal AI systems such that they would be rewarded for taking the ambiguity, vagueness, and complexity of law into account, like skilled legal reasoning does. This framework serves as one of the most pressing issues facing legal AIs: overconfident responses to uncertain legal inquiries, even at state-of-the-art human baseline accuracies in top predictions (Koenig et al., 2023).

2. The Inadequacy of Deterministic Benchmarks

2.1 Current Evaluation Paradigms

Existing legal AI benchmarks often depend on a model's accuracy to answer questions with known correct answers. With over 700k large language models on repositories such as HuggingFace, there is a push to establish widely accepted measures of accuracy, robustness, and reasoning on commonly used tasks for benchmarking (along with competitive pressure) in the legal AI community (McIntosh et al., 2025). This mostly means computer-scored bar exam questions, standard legal test questions, or settled cases with agreed-upon outcomes. However, automated metrics are ill-suited to capturing complex legal reasoning (McIntosh et al., 2025), and while such metrics can set useful baseline levels of skill, they are ill-suited for capturing the ambiguity of real-world legal problems. In addition, empirical data demonstrated that variations as small as a change in choice indicator or an added space affect the model's response by at least 5%, indicating that the majority of the model's responses rely on superficial features rather than legal reasoning. Additionally, the benchmarks generally present deterministic questions, which allow the models to express certainty in their responses rather than subtle or interpretative ones, incentivizing model responses to be single answers (McIntosh et al., 2025).

2.2 The Reality of Legal Ambiguity

Several further gradations of vagueness are obvious in legal practice that deterministic benchmarks fail to address, such as statutes using vague language, the need to interpret a text in a specific context, and the fact that jurisdictions can reach opposing conclusions in factually similar cases. With the shifting social context in which a question arises, previous analyzes often lose their relevance, and the resulting period of flux in the law often engenders circuit splits (Eriksson et al., 2025). Such fact-sensitive standards as reasonableness or good faith or substantial similarity, however, cannot be so mechanically analyzed or applied; categorical approaches to them prove elusive. The law provides rich resources for the interpretation of vague or incomplete descriptions of collective human intentions: whole repertoires of variegated applications and generalized precedents with reasons (Nay, 2023). However, benchmarks still cannot distinguish reasoning from pattern-matching because they cannot test whether models understood the interpretive approach of legal reasoning or if they were exploiting surface features of the tests they had trained on (McIntosh et al., 2025). Additionally, benchmarks are unable to evaluate the utility and applicability of models in the presence of legal ambiguity and uncertainty, as they can easily succeed by pattern-matching at this stage, without understanding how the law must be reasoned about (McIntosh et al., 2025; Nay, 2023).

2.3 Consequences of Mismatch

Deterministic benchmarks that incarnate structural legal indeterminacy give rise to predictable failures. In split-authority situations, models report the single interpretation that would be judged correct by an external observer and ignore alternative interpretations. Faced with harder questions, models occasionally refuse to answer, invoking the harmlessness over helpfulness principle to sidestep interpretive complexity (McIntosh et al., 2025). They gloss over statutory tensions that would make a reasonable and conscientious lawyer hedge their answer, misrepresenting their ethical consideration as a failure of interpretive complexity (McIntosh et al., 2025). Lack of qualifiers and indicators of uncertainty in AI writing may mislead legal users into committing malpractice, violating legal ethics, or making bad planned decisions, weakening the value human practitioners derive from AI assistance. This problem is compounded by the mismatch between training objectives and actual practice, given there is more than a 10% chance that transformative AI is developed within 15 years and a roughly 50% chance within 40 years (Nay, 2023). Ill-considered legal AI could thus be deployed at an unprecedented scale in the near future. In light of this uncertainty, benchmarks may inadequately measure reasoning competency and exacerbate the high utility and cognitive capabilities of AI systems (in the applied sense) if used inappropriately as part of the training data (McIntosh et al., 2025). We may need to rethink the evaluation mechanisms to encourage legally appropriate behavior in AI systems. This is necessary so that learning goal specification and interpretation methods can be implemented that adapt to reflect society's changing expectations as encoded in legal regulations (Nay, 2023).

Failure Mode	Manifestation in Legal Context	Measurement Gap	Frequency/Impact
Prompt Sensitivity	Choice indicator format changes alter legal interpretation selection	No standardized prompt robustness testing	5% (approximately) accuracy variation from minor formatting
Pattern Recognition Masking	Model exploits familiar benchmark elements without genuine legal reasoning	Cannot distinguish reasoning from memorization	Artificially inflated scores on repeated benchmark exposure
Evasion Behavior	Refuses complex ambiguous queries to maintain "harmless" profile	Helpfulness vs. harmlessness trade-off unmeasured	Misleading ethical compliance appearance
Training Data Contamination	Benchmark questions inadvertently included in model training corpus	No verification of benchmark-training data separation	Unknown prevalence, validity concerns
Commercial Optimization Pressure	Models tuned specifically for benchmark performance metrics	Gap between benchmark scores and practical utility	Industry ownership drives metric gaming

Table 1: Benchmark Failure Modes in Legal AI Evaluation (McIntosh et al., 2025)

3. Taxonomy of Legal Ambiguity

3.1 Deterministic Questions

Deterministic legal issues are regulated by statutes from only a single source that cannot be interpreted in conflicting ways. In these cases, computational legal systems automatically discover rules and reach conclusions with very high predictive accuracy (Esposito, 2021). These may include bright-line rules, unambiguous statutes, and uniform doctrines with overwhelming jurisdictional consensus, creating practical opportunities for automating document generation, due diligence, and sorting legal information in ways that are fast, inexpensive, accurate, and efficient in ways that would have been previously only practical for human sources (Esposito, 2021). The process includes deterministic tasks such as fixed statutory filing deadlines and legal definitions and legal operations codified in law. Tests like these demonstrate the effectiveness of modern LLMs. Claude 3 Opus had 87.13% accuracy with an F1 score of 0.87 on topic classification tasks and 86.8% on general reasoning tasks (compared to 86.4% for GPT-4 and 83.7% for Gemini 1.0 Ultra) (Sargeant et al., 2025). However, these queries are only a small subset of legally perplexing questions where interpretation is required. Deterministic queries are useful baseline measures of minimum competence but tell us little about how AI performs in real-world legal reasoning tasks that involve legitimate uncertainty and are analogous to the legal reasoning tasks that are central to the autonomy of legal conversation in the modern rule of law (Esposito, 2021; Sargeant et al., 2025).

3.2 Contested Questions

Contested questions arise when a case is considered a legitimate interpretive disagreement between the courts, scholars, or practitioners. The problem in these cases is not that the algorithm explains too little, but that it explains too much and too precisely. These cases expand on circuit splits, in which different appellate courts adopt conflicting interpretations of federal statutes or federal constitutional clauses, to include other areas of law, such as doctrines in emergent areas where the law is unsettled but moving. Contested questions also fall into the same analogical reasoning processes: classifying relevant precedents into analogous or distinguishable leads to different decisions. Topic modeling on judicial corpora also reveals contested questions: the BERTopic model identified topics of US Supreme Court decisions with 84.6% semantic similarity as judged by experts, and the LSA model identified topics of European Court of Human Rights decisions with 43.1% semantic similarity and a 91.4% area under the receiver operating characteristic curve. This evidence shows that computational systems may differ in the interpretive disagreement they detect, depending on the domain. This raises questions for lawyers about mapping interpretive disagreements, the strongest arguments on each side, and jurisdiction. Should these questions be answered by algorithmic processes, their lack of transparency may weaken contestability, since systemic or confirmation bias may reproduce or compound the existing bias in the machine's training data (Esposito, 2021). Effectively answering them requires recognizing the deficiencies of the machine and is thus a meta-communicative, co-constructive process between machine and interlocutor: the machine is involved in the meta-communication, the interlocutor in the process and data of the machine, and the interlocutor interprets rather than understands the machine's process (Sargeant et al., 2025).

3.3 Indeterminate Questions

Indeterminate questions, such as policy-laden decisions or fact-sensitive thresholds, do not have determinate legal answers. Thus, artificial communication instead of artificial intelligence may be needed because the objective of explanation should not be to make algorithmic procedures intelligible but to reformulate them in a way understandable to legal actors (Esposito, 2021). An example of such situations is the open-textured legal standard of the "reasonable" person standard or the "fair use" standard in copyright law. In humans, the accuracy of their explanations is not dependent on the transparency of neural paths or thought

processes. For algorithms, explanations are not limited to disclosing the machine process and may just be a contextually appropriate response (Esposito, 2021). Indeterminate questions arise in new fact patterns for which existing legal categories are poorly suited, requiring analogical extension or doctrinal innovation predicated on interpretive discretion, not mechanical application. Indeterminate questions include those whose answers depend on factually controversial predicates and those whose answers depend on predictions about the future, where those predictions are uncertain. Large language models have varied abilities to answer indeterminate questions. In legal reading comprehension tests, Claude 3 Opus scored 83.1%, GPT-4 scored 80.9%, and Gemini 1.0 Ultra scored 82.4%. However, indeterminate questions cannot be answered by legal research alone but require a combination of legal principles with contextual judgment, policy considerations, and probabilistic reasoning in light of factual uncertainties (Sargeant et al., 2025). The issue thus becomes whether machines are able to provide satisfactory explanations in response to the different requests of their interlocutors, given that the requests for clarification are always different and that the dump approach often, but not always, varies with mentalizing dialogue to control without transparency (Esposito, 2021).

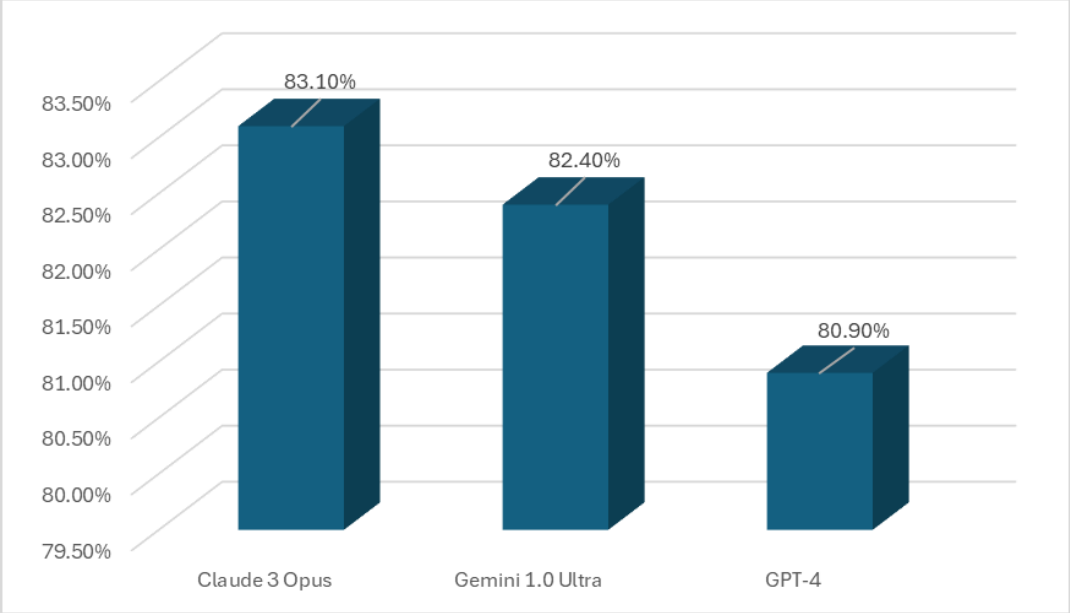

Figure 1. Performance Score of Large Language Models (Sargeant et al., 2025)

3.4 Spectrum Representation

Legal questions do not fall into dispositive or indeterminate categories; instead, they exist along a continuum of determinacy and indeterminacy, as both the indeterminacy of legal language and the indeterminacy of legal interpretation are intrinsic to legal practice in any legal system (Esposito, 2021). Computational analysis, like legal reasoning, is context sensitive: topic modeling, for example, fared better on domain name disputes than ECHR cases, with 91.4 AUROC (Area under Receiver Operating Characteristic Curve) for LSA (Latent Semantic Analysis) and 84.6 AUROC for topic modeling semantic correspondence on Supreme Court legal opinions (Sargeant et al., 2025). While historical doctrine can be settled, its application may remain open, creating a need for meta-communication, communication about communication, between the algorithm and its users beyond algorithmic transparency alone (Esposito, 2021). This framework captures the subtypes of ambiguity discussed above by attaching a series of annotations reflecting these subtypes for any given question. These include interpretive disagreement among courts, vagueness in the statutory text, factual sensitivity of outcomes, policy concerns requiring value judgments, and rapid change in doctrine. This multidimensionality recognizes that the autonomy of legal communication, which is essential to the embedded rule of law, is more about interpretive openness and contestability of legal decisions than about algorithmic precision and deterministic explanations (Esposito, 2021). The danger is that legal decision systems may lose autonomy when they focus on deterministic explanations rather than being part of the dialogic process of legal interpretation, where controlled ambiguity allows for legitimate disagreement and flexible application of legal principles (Esposito, 2021; Sargeant et al., 2025).

4. Constructing an Ambiguity-Aware Benchmark Dataset

4.1 Source Selection

Ambiguity-aware dataset construction requires consideration of the realities of legal-query annotation, such as the presence of ambiguous linguistic forms, domain shifts, and a lack of domain experts to serve as annotators. Contested legal questions can be extracted from databases of circuit splits, diverging state supreme court opinions, or doctrinally unsettled issues. Ambiguous questions differ in their underlying ground patterns: 79% of ambiguous question/answer pairs ground differently, but only 64% of

unambiguous question/answer pairs ground the same. Indeterminate question types arise in cases involving fact-sensitive standards, first impression questions, and policy determinations. Linguistic analyzes can disambiguate ambiguities in question transcriptions in various cases. For example, unambiguous questions are over three times as likely to include plural nouns as ambiguous questions. As they differ in syntax, these can be used as syntactic diagnostics for source detection. Sources need to be sampled across legal domains (constitutional, statutory, common law, and regulatory) to enable controlled variation in annotation difficulty (Chen et al., 2025). Benchmarks should comprise recently decided cases with fluid doctrine and longstanding areas of reasonable disagreement. On average, open-sourced models perform 12.59% worse than proprietary multi-modal large language models on reasoning across these cases. More work is needed to examine how performance varies across model types and set representative difficulty levels based on these benchmarks (Wang et al., 2025).

Category	Metric/Observation	Value
Grounding Patterns	Ambiguous question/answer pairs ground differently	79%
Linguistic Diagnostics	Plural noun frequency in unambiguous vs. ambiguous questions	3× higher
Model Performance	Open-source models perform worse than proprietary models on legal reasoning	12.59% gap
Legal Domain Coverage	Required source types	Constitutional, statutory, common law, regulatory
Question Types	Indeterminate sources	Fact-sensitive standards, first impression cases, policy determinations
Benchmark Composition	Temporal coverage	Recent cases (fluid doctrine) + longstanding disagreements

Table 2: Legal Query Ambiguity Characteristics and Source Selection Criteria [8, 10]

4.2 Annotation Protocol

Questions must be annotated in detail using structured frameworks because all natural language processing systems are reliant on high-quality annotated datasets for training and evaluation and because model generalizability and robustness depend on the consistency, variability, and semantic coherence of the annotated data (Onan et al., 2025). For contested questions, annotators must identify alternative positions, provide authority for each position, and note jurisdictional differences. Multi-agent annotation enables coordinated agent-based collaboration that combines reasoning and task adaptiveness by dividing labor across autonomous agents with complementary capabilities (Onan et al., 2025). Annotations should clarify whether a question is purely legal interpretation or depends on factual characterization. Ambiguous questions have a 79% divergence in question-answer groundings. Ambiguity needs to be explicitly annotated (Chen et al., 2025). Likewise, for indeterminate questions, the policy, value judgments, or future assessments that render the answer indeterminate need to be annotated. Rather than providing a single "correct" label, for each case the dataset provides an "interpretation space" of possible legal interpretations: which interpretations are legally possible, the legal reasoning for them, and their frequency across different jurisdictions. Based on ensemble methods, the dataset introduces a controlled amount of uncertainty in the candidate labels to increase the chance of sampling a high-quality candidate and to surface underlying disagreements that can be resolved post hoc by verification and arbitration. Ambiguity resolution modules have been shown to successfully resolve over 80% of contradictions into a unified single resolution (Onan et al., 2025). They can be sensitive to prompt wording, decoding randomness, and domain mismatch, so careful and principled divergence-annotation protocols that track such variances rather than enforcing a single 'truth' are critical (Onan et al., 2025; Chen et al., 2025).

4.3 Diversity and Balance

All levels of ambiguity should be included in the dataset to enable comparison of performance. For example, the average performance gap of 12.59% exists between open-sourced and proprietary models, suggesting that these models exhibit different sensitivities to particular question features (Wang et al., 2025). The inclusion of deterministic questions establishes a baseline of competence and prevents hedging. Because questions with clear determinate properties have deterministic grammatical features (e.g., being more than three times more likely to contain plural nouns), these samples can be stratified (Chen et al., 2025). Contested questions should vary in circuit, state, degree of doctrinal development, and grounds of the analogies relied upon. For ambiguous questions, 79% of sampled ambiguous questions should have the characteristic divergent grounding pattern of question v. answer formulations in validated ambiguous contexts (Chen et al., 2025). Indeterminate questions should also differ depending on whether their indeterminacy is caused by vague statutory language, factually driven standards or policy-driven discretion. Given that only

64% of unambiguous questions have matching groundings, this finding indicates that the remaining 36% must fall in the intermediate region (Chen et al., 2025). Geographic and temporal diversity ensure coverage of both long-standing doctrinal disputes and recent doctrinal ambiguities; subject matter diversity also serves to defend against overfitting to a single area of law. This is especially important, given that performance on models can vary across domains. By controlling for this variation in the training data, the system can genuinely be distinguished by recognizing ambiguity, and this being simply an artifact of the training distribution (Onan et al., 2025). The framework should include multiple ensemble generators to compare differences in how models of different architecture or training regimes approach the same legal case and to highlight biases in how ambiguity is recognized (Onan et al., 2025).

4.4 Validation and Refinement

Domain experts can help validate the classes of ambiguous tokens as well as the spaces of interpretation through the use of multi-agent systems that incentivize agents to converge on annotation decisions, such as the law in low-resource or culturally sensitive settings where conventional annotation is not possible. Throughout this process, multiple annotators can glean inter-rater reliability, resolve conflicts through structured debate, and introduce ambiguity resolution strategies that can achieve similar results in over 80% of cases (Onan et al., 2025). For a valid test, validators should confirm that ambiguously labeled questions are 79% divergent in grounding distribution (as ambiguous questions are) rather than 64% convergent in grounding context (as unambiguous questions are) (Chen et al., 2025). Furthermore, interpretation spaces required to answer questions should ideally reflect mainstream rather than niche views. The post-arbitration topic classification and sentiment analysis correction rates of 8.3% and 12.1% indicate that while most labels were well-formed, low-level normalizations to match formatting and remove minor linguistic noise were indeed required (Onan et al., 2025). Pilot testing with human lawyers and models can identify ambiguous question classifications that are problematic or identified inconsistently and provide useful feedback on annotation quality. For example, does the 12.59% gap in performance between open-source and proprietary models above reflect differences in model performance or disambiguation, or is it due to the construction of the benchmark (Wang et al., 2025)? Iterative refinement may then be employed to ensure the benchmark is reflective of the type of ambiguity encountered in legal practice while observing that removing these modules degrades model performance by no more than 1 point but is required for format consistency and strict taxonomy adherence in deployed use cases (Onan et al., 2025). The general approach of generating, arbitrating over, and refining the generated output shows that the quality of annotated data stems from filtering and correction, not diverse generation (Chen et al., 2025).

5. Enhanced AI Architectures and Training Methodologies

5.1 Ambiguity Detection Mechanisms

Standard legal AI systems also need to be extended with explicit ambiguity detection components, because vague and ambiguous wording of requirements and legal texts may lead to misinterpretations, incorrect implementations, and wasteful expenditures (Izhar et al., 2025). This may be done, for example, using retrieval-based mechanisms that search through conflicting authorities with regard to legal questions. This is because no one method is good enough to always cope with the existing ambiguity in requirements (Izhar et al., 2025). Examples of interpretive ambiguity include systems trained to recognize indicators of interpretive disagreement (such as "courts have split," "the majority view," and "however, some jurisdictions") and systems trained to recognize structural indicators such as contradictory precedent based on the same facts or vague words (for example, "reasonable," "substantial," and "appropriate"). Hybrid rule-based and machine learning systems have attained 97% accuracy, 97% precision, 89% recall, and 92% F1 in the domain of interpretation ambiguity with a methodology similar to that successfully applied in software requirements analysis; these outperformed customary rule-based systems. For example, TF-IDF vectorization of text and Random Forest classifiers can recognize complex signals of ambiguity that classic rule-based systems obscure (Izhar et al., 2025).

5.2 Comparative Reasoning Performance Across LLM Architectures

Modern large language models differ widely in their abilities to reason about legal documents, including a distinction between raw factual reasoning and intellectual fact-based reasoning. Evaluating five leading generative artificial intelligence systems, we show that various architectural choices explain systematic differences across models (McIntosh et al., 2024). The model with the best reasoning capabilities is OpenAI ChatGPT-4 with a reasoning score of 96.5 for raw facts and 93.5 for intellectual facts (i.e., a three-point degradation in reasoning performance from raw to intellectual). The model with moderate reasoning capabilities is OpenAI ChatGPT-3.5, with a slight degradation, from 89 to 81.5. Google Bard has a score of 89 for reasoning with raw facts but a score of 62.5 for reasoning with intellectual facts. With Perplexity AI, the raw and intellectual fact-reasoning abilities are 92 and 44, respectively, indicating a 48-point drop that reflects a principled fundamental architectural limit. TruthGPT performed the worst of all these models, receiving a score of 60 on the raw facts and 14.5 on the intellectual facts, which represents a huge decline (McIntosh et al., 2024).

These results reveal several important requirements for the architecture of legal AI, such as the amount of performance drop when moving from factual to intellectual reasoning for most models (McIntosh et al., 2024). Legal AI architectures need to detect when

reasoning moves from factual to intellectual cases and to calibrate themselves accordingly, acknowledging the uncertainty of reasoning in intellectual cases and avoiding spurious certainties.

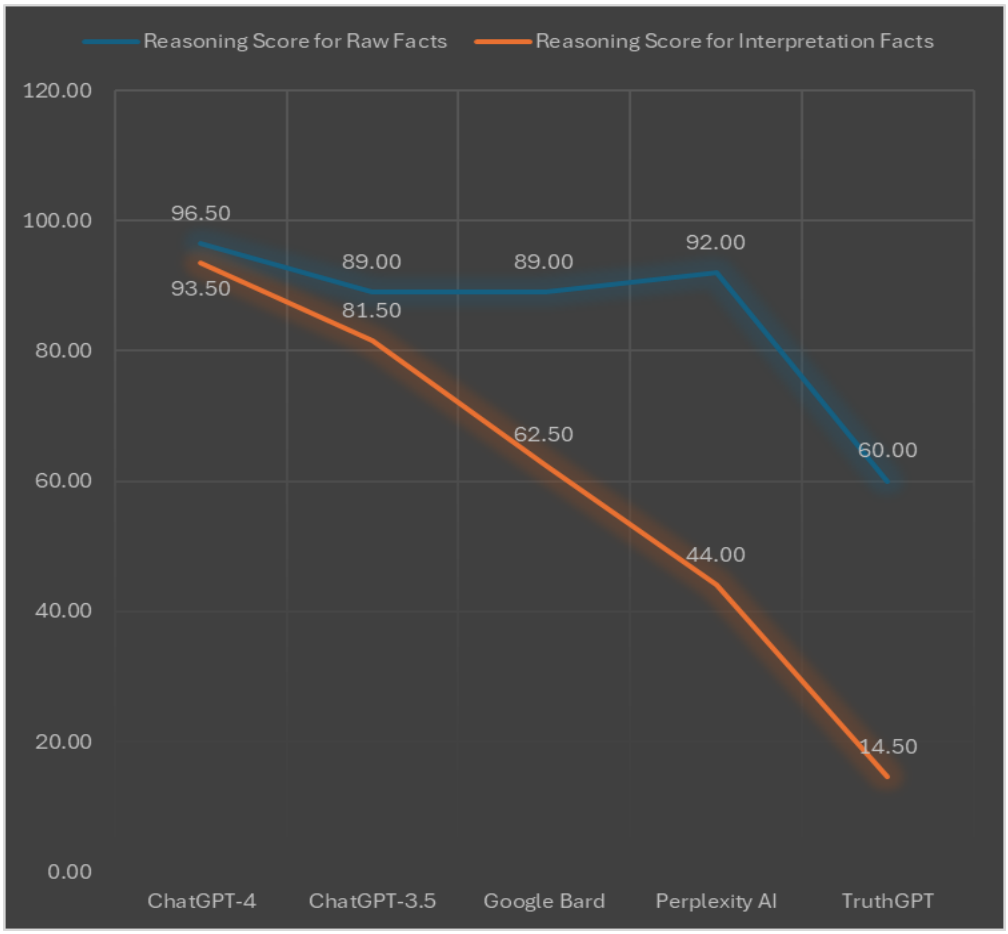

Figure 2: GPT Models with Reasoning Capabilities (McIntosh et al., 2024)

5.3 Interpretation Enumeration

The systems must also be able to disambiguate these interpretations, for example by utilizing rule-based or machine learning approaches for identifying vague or uncertain language to lower the risk of misinterpretation (Izhar et al., 2025). Extensions to architectures can also include structured generation, which produces multiple representations before a response is generated. Methods that have achieved 97% accuracy on classification tasks include clustering analysis to find a feature representation of the text for identifying similarities, as well as supervised methods that use sentiment analysis to find relationships between ambiguity and text tone (Izhar et al., 2025). Retrieval-augmented generation can be prompted to enumerate positions as majority or minority, extract key rationales in support, and cover the space of legal interpretation better than systems validated on 1553 requirements; it can also help with understanding ambiguous patterns and stakeholder views. It can generate arguments for conflicting sides to motivate the entire interpretation space. It must avoid the limit, however, that legal documents such as court filings, judicial judgments, statutes, treaties, contracts, and formal legal correspondence encode rules, rights, and duties from which legal practitioners, judges, regulators, and academics derive legal case analysis, legislative interpretation, and compliance verification. These documents can show how legal terminology and textual conventions can impact enumeration efforts to discover legal reasoning shortcomings. Chain-of-thought prompting allows models to consider counterarguments, read the relevant legal authority, consider jurisdiction differences, and draw conclusions. Alternatively, explaining the success of predicting case outcomes based solely on judges' surnames might require the enumeration systems for legal interpretation to explicitly counter the data bias on judicial prediction and limit the propagation of systematic bias through their approaches (Izhar et al., 2025).

5.4 Uncertainty Caliberation

Furthermore, due to interpretational ambiguity, Legal AI must also learn to hedge uncertainty by providing frameworks that clarify the decision-making process while also protecting decisions through flexible, interpretable solutions (Izhar et al., 2025). For legal

professionals, a suitable approach is a training dataset with both correct answers and hedging annotations. The results show that using multiple dimensions with end-to-end models achieves a precision rate of 97%, a recall rate of 89%, and an F1 score of 92%, reflecting high performance and interpretive uncertainty (Izhar et al., 2025). Prompt engineering for hedged language (e.g., "the majority of courts", "under the prevailing view", "this issue remains contested") can be part of approaches considering issues of precision and coverage of classification (Izhar et al., 2025). Fine-tuning on calibration metrics can also punish hallucination by favoring uncertainty when faced with contested issues and rewarding confidence when dealing with deterministic queries. This is particularly concerning because large language models hallucinate legal content on more complex tasks (McIntosh et al., 2024). One challenge is scaling up uncertainty calibration modeling to stronger and more interpretable systems that are more explainable to people, particularly in terms of reasoning and language of law, which are complex and context dependent. In addition to these issues, with respect to AI cases in the legal field, there may be difficulties debiasing the system when 65% accuracy on the basis of predictions based on surnames suggests that calibrated uncertainty must recognize interpretive ambiguity variable from bias in the training data (Izhar et al., 2025).

5.5 Comparative Training Objectives

Instead of optimizing for unquestionable correctness, outputs that acknowledge and account for ambiguity shall be incentivized because, as complex, distributed software systems and legal frameworks are built, ambiguity results in more costly human miscommunication and software development and adjudication failures (Izhar et al., 2025). Rewards should incentivize getting into the habit of mentioning minority positions and being careful about claims. These should be informed by a clustering analysis to reveal ambiguity and diversity among stakeholders in the interpretation space (Izhar et al., 2025). Contrastive learning could allow systems to infer whether a context is deterministic or ambiguous and to adapt its response style accordingly, addressing the shortcomings of existing approaches on the three aspects of scalability, accuracy, and interpretability (Izhar et al., 2025). Human lawyers could also provide reinforcement learning by grading the answers based on the coverage of interpretation and the appropriateness of expressed certainty, rather than simple correctness. Evaluation would have to compare correctness against authority by accessing legal texts through official court websites, statutory databases, and proprietary commercial legal research tools (McIntosh et al., 2024). The hallucination rate of 58% legal content in some cases limits comparative training since its objective would penalize the invention of authority and reward the model for acknowledging uncertainty (McIntosh et al., 2024).

6. Conclusion

The proposed Ambiguity-Aware Evaluation Framework would provide a substantial step forward from deterministic accuracy benchmarks for evaluating legal AI, shifting emphasis from deterministic measures to indeterminate questions. In particular, it would involve not just a typology for distinguishing between deterministic and indeterminate questions, but also appropriate measures for calibrating uncertainty. Augmented models that support ambiguity detection, interpretation enumeration, and qualified reasoning better reflect professional legal discourse. While such models do not improve accuracy on objectively answerable questions, they outperform mainstream models on disputed and indeterminate questions, better matching tasks of interest in more advanced legal practice. On this basis, it is suggested that these systems could adopt an epistemic humility perspective, especially in situations where there is no judicial consensus. In the future, these systems will be improved by increasing the representation of domains in benchmarks, calibrating uncertainty estimates, and validating performance through deployment in legal practice, eventually overlapping with the careful and hedged reasoning of high-quality professional legal advice.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Koenig, N., Tonidandel, S., Thompson, I., Albritton, B., Koohifar, F., Yankov, G., Speer, A., Hardy, J. H., III, Gibson, C., Frost, C., Liu, M., McNeney, D., Capman, J., Lowery, S., Kitching, M., Nimbkar, A., Boyce, A., Sun, T., Guo, F., Min, H., Zhang, B., Lebanoff, L., Phillips, H., & Newton, C. (2023). Improving measurement and prediction in personnel selection through the application of machine learning. Personnel Psychology. https://onlinelibrary.wiley.com/doi/10.1111/peps.12608

[2] Savelka, J., Holzenberger, N., & Van Durme, B. (2023). Can GPT-4 support analysis of textual data in tasks requiring highly specialized domain expertise? Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law. https://dl.acm.org/doi/10.1145/3594536.3595163

[3] McIntosh, T. R., Susnjak, T., Arachchilage, N., Liu, T., Xu, D., Watters, P., & Halgamuge, M. N. (2025). Inadequacies of large language model benchmarks in the era of generative artificial intelligence. IEEE Transactions on Artificial Intelligence. https://arxiv.org/pdf/2402.09880

[4] Eriksson, M., Purificato, E., Noroozian, A., Vinagre, J., Chaslot, G., Gomez, E., & Fernandez-Llorca, D. (2025). Can we trust AI benchmarks? An interdisciplinary review of current issues in AI evaluation. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8(1), 850–864. https://arxiv.org/abs/2502.06559

[5] Nay, J. J. (2023). Law informs code: A legal informatics approach to aligning artificial intelligence with humans. Northwestern Journal of Technology and Intellectual Property. http://arxiv.org/pdf/2209.13020

[6] Esposito, E. (2021). Transparency versus explanation: The role of ambiguity in legal AI. Journal of Cross-disciplinary Research in Computational Law. https://journalcrcl.org/crcl/article/download/10/8

[7] Sargeant, H., Izzidien, A., & Steffek, F. (2025). Topic classification of case law using a large language model and a new taxonomy for UK law: AI insights into summary judgment. Artificial Intelligence and Law, 1–49. https://link.springer.com/content/pdf/10.1007/s10506-025-09434-0.pdf

[8] Wang, R., Song, S., Ding, L., Gong, M., Iwasawa, Y., Matsuo, Y., & Guo, J. (2025). MMA: Benchmarking multi-modal large language model in ambiguity contexts. ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models. https://openreview.net/pdf?id=kY8k5iHMmu

[9] Onan, A., Nasution, A. H., & Celikten, T. (2025). Toward reliable annotation in low-resource NLP: A mixture of agents framework and multi-LLM benchmarking. IEEE Access, 13, 211620–211644. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=11299523

[10] Chen, C., Tseng, Y.-Y., Li, Z., Venkatesh, A., & Gurari, D. (2025). Acknowledging focus ambiguity in visual questions. Proceedings of the IEEE/CVF International Conference on Computer Vision, 1228–1238. https://openaccess.thecvf.com/content/ICCV2025/papers/Chen_Acknowledging_Focus_Ambiguity_in_Visual_Questions_ICCV_2025_paper.pdf

[11] Izhar, R., Bhatti, S. N., & Alharthi, S. A. (2025). Bridging precision and complexity: A novel machine learning approach for ambiguity detection in software requirements. IEEE Access. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10843220

[12] McIntosh, T. R., Liu, T., Susnjak, T., Watters, P., & Halgamuge, M. N. (2024). A reasoning and value alignment test to assess advanced GPT reasoning. ACM Transactions on Interactive Intelligent Systems, 14(3), 1–37. https://dl.acm.org/doi/pdf/10.1145/3670691