Bridging Legal Interpretation and Formal Logic: Faithfulness, Assumption, and the Future of AI Legal Reasoning

Olivia Peiyu Wang
University of California, Santa Cruz
pwang95@ucsc.edu

Leilani H. Gilpin
University of California, Santa Cruz
lgilpin@ucsc.edu

Abstract

The growing adoption of large language models in legal practice brings both significant promise and serious risk. Legal professionals stand to benefit from AI that can reason over contracts, draft documents, and analyze sources at scale, yet the high-stakes nature of legal work demands a level of rigor that current AI systems do not provide. The central problem is not simply that LLMs hallucinate facts and references; it is that they systematically draw inferences that go beyond what the source text actually supports, presenting assumption-laden conclusions as if they were logically grounded. This proposal presents a neuro-symbolic approach to legal AI that combines the expressive power of large language models with the rigor of formal verification, aiming to make AI-assisted legal reasoning both capable and trustworthy, thus reducing the burden of manual verification without sacrificing the accountability that legal practice demands.

1 Problem Statement: Two Kinds of Correctness

Legal professionals and AI researchers share the goal of deriving trustworthy, verifiable reasoning over legal text, but approach it through fundamentally different frameworks. Legal interpretation draws on background knowledge and contextual inference, while formal logic demands that all inferences be explicitly grounded in the text.

These frameworks are not different levels of rigor, but different modes of reasoning. A legally sound conclusion may be formally invalid because that norm is nowhere stated in the contract. This gap is largely invisible in legal AI research because most systems either mimic legal interpretation through language model training or enforce formal validity through symbolic methods, without acknowledging that the two regularly diverge. We argue that making this gap explicit, measurable, and addressable is one of the most important open problems in legal AI.

2 Research Contribution: Measuring the Gap

Dataset and re-annotation. We investigate this gap using ContractNLI (Koreeda and Manning, 2021), one of the few benchmarks grounded in authentic contract language. Its original labels, produced by legally trained annotators, largely reflect legal interpretation. We do not dispute this. We re-annotate the same examples under a strict formal definition - a hypothesis H is entailed by premise P if and only if $P \wedge \neg H$ is unsatisfiable, contradicted if and only if $P \wedge H$ is unsatisfiable, and neutral otherwise.

The result is striking - there is a substantial proportion of label shifts, primarily from ENTAILMENT to NEUTRAL. These are not errors. There are cases where the original conclusion depends on background legal knowledge or contextual assumptions reasonable for a lawyer to invoke but absent from the text. This predominant ENTAILMENT→NEUTRAL transition reveals a systematic gap between pragmatic legal interpretation and strict formal entailment. We further construct minimal pairs - for each divergent case, a minimally modified hypothesis that becomes formally entailed by supplying the missing assumption explicitly. This approach transforms an opaque gap into a tractable, analyzable object.

Experimental findings. We evaluate three paradigms across five LLMs (GPT (Agarwal et al., 2025), Claude (Anthropic, 2026), LLaMA (AI@Meta, 2024), DeepSeek (Liu et al., 2025), Qwen (Yang et al., 2024)): pure LLM classification, LLM reasoning over formal logical representations, and a neuro-symbolic pipeline combining LLM formalization with an SMT solver (de Moura and Bjørner, 2008). Formal structure improves

accuracy, but **accuracy does not imply faithful reasoning**. High-performing models succeed by mimicking legal interpretation, including its implicit assumptions, rather than by reasoning formally. The SMT pipeline is more conservative, returning neutral classification whenever explicit grounding is lacking, and surfacing the gap rather than papering over it.

We identify three recurring failure modes

- **Assumption Injection:** the reasoning silently bridges gaps with unstated inferences.
- **Scope Laundering:** the reasoning presents informal conclusions as formally grounded.
- **Implicit Constraint Blindness:** the reasoning overlooks constraints present in formal representations

The dominant error across all models is NEUTRAL $\rightarrow$ ENTAILMENT misclassification, reflecting systematic assumption injection. The ENTAILMENT $\leftrightarrow$ CONTRADICTION confusions are rare, which indicates that the challenge is insufficient grounding, not logical inconsistency.

Minimal axiom framework. Furthermore, the NEUTRAL classification does not necessarily need to be a dead end. Rather than stopping at a neutral classification, the system computes the *minimal set of additional axioms* sufficient to shift the classification to ENTAILMENT or CONTRADICTION, and presents them to a legal reviewer with a targeted question - does standard contract law or the contractual context implicitly establish this assumption? A lawyer answering *yes* validates the implicit norm that formal logic cannot capture alone; a lawyer answering *no* confirms the case is genuinely underspecified. In both cases, legal expertise is applied precisely where formal methods reach their limit. The complexity of the minimal axiom set is also diagnostic - many or complex axioms signal genuine interpretive difficulty, while few and simple axioms suggest confident automated classification.

3 Research Agenda: Bridging the Gap

The legal interpretation-formal logic gap is not unique to contract entailment. It arises wherever AI systems cite sources to support legal claims, which is increasingly common in LLM-assisted drafting, regulatory analysis, and litigation support (Magesh et al., 2025; Freifeld and Scarcella, 2026). Prior work addresses *fabricated citations* such as references that do not exist, a problem now largely tractable through retrieval verification (Agrawal et al., 2024; **?**). The more difficult and consequential problem is *stance misrepresentation* where the source exists and is retrieved correctly, but the claim overstates, understates, or mischaracterizes what the source says (United States Court of Appeals for the Sixth Circuit, 2026). A model trained to reason like a lawyer will routinely infer more from a cited source than it strictly supports, because that is what legal interpretation does. Between 50 and 80 percent of LLM responses in legal and medical domains are not fully supported by their cited sources (Agrawal et al., 2024), yet no formal detection framework exists for this failure mode.

We propose two contributions to address this -

- **A Benchmark Dataset:** The benchmark dataset should consist of LLM-generated legal and academic text annotated for stance misrepresentation at the claim level, using the three-way framework and minimal pair methodology from our research contribution.
- **Solver-in-the-Loop Training:** Rather than using LLM judges or human preferences as feedback, we will use a formal verification tool as a reward signal, teaching the model to distinguish formally supportable inferences from assumption-laden ones. When the solver flags an insufficiently grounded claim, it also computes the minimal axioms required to ground it, feeding directly into targeted human review at exactly the points where legal interpretation and formal grounding diverge.

Vision. The long-term goal is AI legal reasoning that operates transparently across both frameworks. It should be capable of legal interpretation when appropriate, formal grounding when required, and able to surface the minimal assumptions connecting the two for targeted legal review (Hildebrandt, 2018; Francesconi and Governatori, 2023). We argue that understanding where and why current legal AI systems break is not a limitation but the foundation of this agenda. Only by honestly characterizing failure modes can we identify where AI assistance can be responsibly applied and build systems that proactively surface interpretive uncertainty rather than asking lawyers to verify conclusions after the fact (Dixon Jr, 2025).

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Kalai. 2024. Do language models know when they're hallucinating references? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 912–928.

AI@Meta. 2024. Llama 3 model card.

Anthropic. 2026. Introducing claude sonnet 4.6. Accessed: 2026.

Leonardo Mendonça de Moura and Nikolaj S Bjørner. 2008. Proofs and refutations, and z3. In *LPAR Workshops*, volume 418, pages 123–132. Doha, Qatar.

Judge Herbert B Dixon Jr. 2025. Guidelines for judicial officers: Responsible use of artificial intelligence. *The Judges' Journal*, 64(2):36–38.

Enrico Francesconi and Guido Governatori. 2023. Patterns for legal compliance checking in a decidable framework of linked open data: E. francesconi, g. governatori. *Artificial Intelligence and Law*, 31(3):445–464.

Karen Freifeld and Mike Scarcella. 2026. Sullivan & cromwell law firm apologizes for AI hallucinations in court filing. https://www.reuters.com/legal/litigation/sullivan-cromwell-law-firm-apologizes-ai-hallucinations-court-filing-2026-04-21/. Reuters. Accessed: 2026-04-29.

Mireille Hildebrandt. 2018. Law as computation in the era of artificial legal intelligence: Speaking law to the power of statistics. *University of Toronto Law Journal*, 68(supplement 1):12–35.

Yuta Koreeda and Christopher D Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

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

United States Court of Appeals for the Sixth Circuit. 2026. United states v. john farris, no. 25-5623 (6th cir. 2026). https://law.justia.com/cases/federal/appellate-courts/ca6/25-5623/25-5623-2026-04-03.html. Decided April 3, 2026.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.