Bridging Legal Interpretation and Formal Logic

Faithfulness, Assumption, and the Future of AI Legal Reasoning

Olivia Wang

Advisor: Prof. Leilani H. Gilpin

Email:

pwang95@ucsc.edu

AIEA Lab, UCSC

What is Legal AI

Why Legal AI

- Legal field can benefit from the advances in AI significantly
  - Due diligence is really time consuming
  - Legal professionals are swamped with legal documents
  - AI have the capability to process a large amount of documents with ease

How it is going

Sullivan & Cromwell law firm apologizes for AI 'hallucinations' in court filing

By **Karen Freifeld** and **Mike Scarcella**

April 21, 2026 9:50 AM PDT · Updated April 21, 2026

US. V. Farris

Howe filed a timely response. In it, Howe admits that he used artificial intelligence to prepare both briefs he filed. According to Howe, he directed an unnamed "staff" member to upload district court documents to Westlaw's CoCounsel program to create a first draft of the principal brief. Howe Show Cause Response at 2. He then worked in that same file for six hours to supplement the draft produced by artificial intelligence. Howe notes that he repeated that same process for the reply brief.

By way of attempted explanation, Howe claims that this appeal was his first time utilizing Westlaw CoCounsel "in this way for a Court of Appeals brief." *Id.* And he says that he was otherwise unfamiliar with the program. Howe's response states that his law office first acquired Westlaw CoCounsel in August 2025—after the district court proceedings

Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools

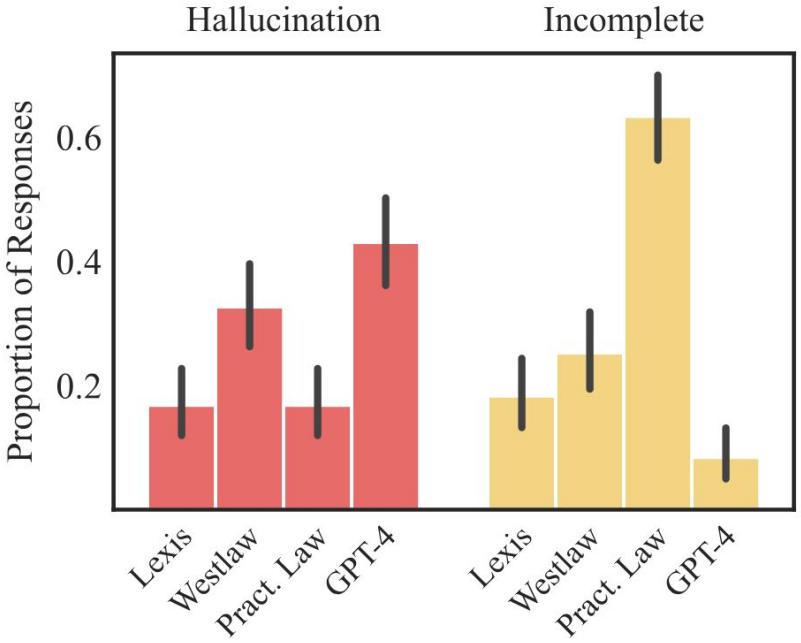


Figure 1: Comparison of hallucinated and incomplete answers across generative legal research tools. Hallucinated responses are those that include false statements or falsely assert a source supports a statement. Incomplete responses are those that fail to either address the user's query or provide proper citations for factual claims.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of empirical legal studies*, *22*(2), 216-242.

The Status Quo

- AI generates hallucinated references, stances, etc.
- Judges overwhelmed with AI generated filings
- Sanctions, fines, and motion dismissals due to AI hallucinations
- Big Law Firms sustained reputation damages from AI hallucinations
- Senior lawyers held accountable for their subordinates' mistakes with AI tools
- Bar Associations issuing guidance on the use of AI

Why Current Legal AI Failed

- GenAI are probability based neural networks that are nontransparent, unreliable, and untrustworthy in high-stakes domains
- Retrieval-Augmented Generation (RAG) helps but not much because the structure of legal systems is not captured
- Scaling alone will not solve the problem
- “Practicing law is a form of art, you learn by practicing”

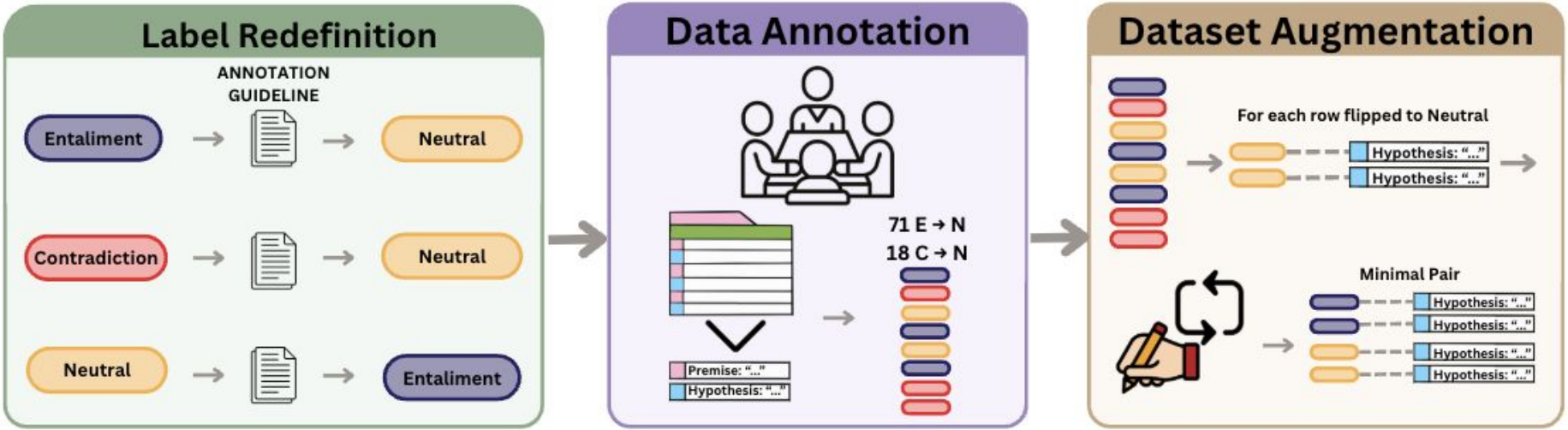
Our Work Shows

1. Adding formal structure improves LLM performance

2. Improved accuracy does not imply faithful reasoning

3. Assumption ambiguity is a core challenge

Dataset Construction



Dataset Construction (Cont.)

Entailment : The premise logically entails the hypothesis if and only if every truth assignment that satisfies the premises also satisfies the hypothesis.

Premise: “Nothing in this Agreement is to be construed as granting the Recipient any right whatsoever with respect to the Confidential Information,”
Hypothesis: “Agreement shall not grant Receiving Party any right to Confidential Information”

Dataset Construction (Cont.)

Contradiction: A logical state where the premises and hypothesis together lead to an impossible outcome

Premise: “Upon request, or if either party elects not to pursue any further business undertaking with the other, Recipient shall promptly return all tangible information, including any and all copies or partial copies thereof and thereupon confirm destruction of all information held electronically.”

Hypothesis: “Receiving Party may retain some Confidential Information even after the return or destruction of Confidential Information.”

Dataset Construction (Cont.)

Neutral (insufficient grounding):

Premise: “The Recipient shall use the Confidential Information solely for the purpose for which it was disclosed,”

Hypothesis: “Receiving Party shall not use any Confidential Information for any purpose other than the purposes stated in the Agreement”

Dataset Construction (Cont.)

Neutral (irrelevance):

Premise: “Retention of confidential information….”

Hypothesis: “The governance of third-party disclosure…”

Impacts of Re-annotation

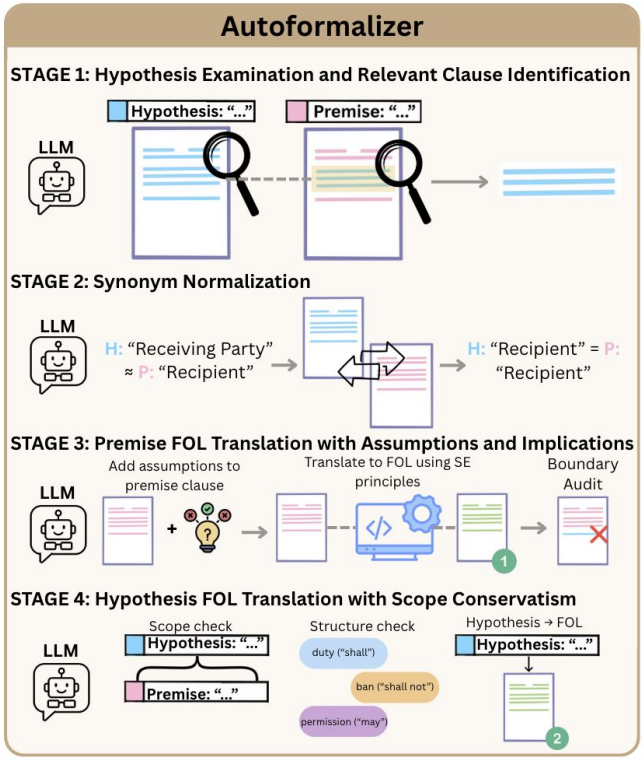
Label Transition	Count
Entailment → Neutral	71
Contradiction → Neutral	18
Neutral → Entailment	14
Neutral → Contradiction	4
Entailment → Contradiction	1
Contradiction → Entailment	1
Total Entailment	153
Total Contradiction	52
Total Neutral	295

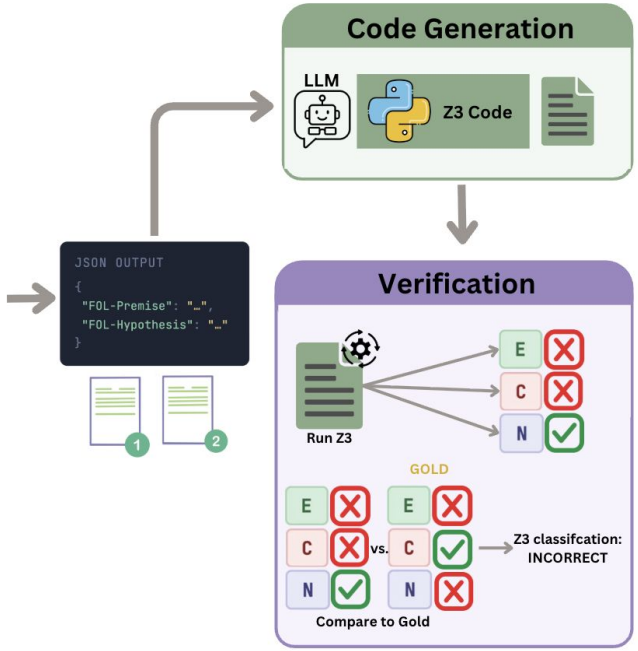
Inter-annotation Agreement Score

Metric	Value
Annotators	Annotator-1 & Annotator-2
Cohen's κ	0.627
Interpretation (Landis & Koch)	Substantial
Percent agreement	81.0%
Percent Disagreements	19.0%

Table 2: Inter-annotator agreement summary.

Neuro-symbolic Pipeline





UNIVERSITY OF CALIFORNIA SANTA CRUZ

Why SMT

1. SMT: A Satisfiability Modulo Theories (SMT) solver is a tool that can formally determine whether a set of formulas in first-order logic is satisfiable with respect to some background theory T.

2. Built-in MCS for legal interpretation
   a. MCS: Minimal Correction Subset
   b. Expose where legal ambiguity lies
   c. Human lawyers as the final decision-makers

3. Industry supported

LLM Experiments Results

Model	Simple	Structured	LLM-Formal Reasoning	SMT-Formal Reasoning
Claude	63.1%	53.2%	83.0%	74.5%
Deepseek	68.4%	62.2%	64.2%	45.3%
GPT	66.0%	58.8%	65.3%	60.3%
Qwen	65.1%	61.5%	69.1%	48.3%
Llama	55.9%	40.0%	58.6%	42.1%

Table 3: Performance Comparison across Classification Approaches. All values report accuracy (%). *Simple*: pure LLM classification using the annotation guideline as the prompt. *Structured*: pure LLM classification with a revised prompt following the multi-stage pipeline outline. *LLM-Formal Reasoning*: LLM-based classification where the premise and hypothesis are autoformalized into First-Order Logic, and the LLM is prompted to produce a classification by running the solver. *SMT-Formal Reasoning*: Z3 solver-based classification where the autoformalized premise and hypothesis are run through the Z3 solver for the classification results

LLM Experiments Results

Model	Error Rate	Invalid (%)
Claude	25.5%	0
DeepSeek	54.7%	23.6%
GPT	39.7%	6.1%
Qwen	51.7%	15.6%
Llama	63.2%	28.7%

Table 5: SMT solving performance across models.

Findings from LLM Experiments Results

1. LLM over-entailment

2. LLM-based Formal Reasoning as intermediate is still problematic
   a. Assumption Injection
   b. Scope Laundering
   c. Implicit Constraint Blindness

3. SMT conservativeness

4. Error asymmetry

1. Structured representations improve performance but LLMs frequently rely on implicit assumptions that are not grounded in the input.

2. Challenge being the absence of a clear boundary between valid inference and unjustified assumption

3. Progress should focus on methods for handling the representation and evaluation of the assumptions

Future of AI Legal Reasoning

1. Minimal Correction Subsets for legal interpretation + Answer Set Programming inspired non-monotonicity

2. Solver-in-the-Loop training

3. Back to Basics: Symbolic-First Reasoning architectures
   a. Model legal systems to guide information retrieval and improve RAG performances
   b. Knowledge-graph to connect the cases that are related

Future of AI Legal Reasoning

- The goal is not to create something that will “replace” human lawyers.

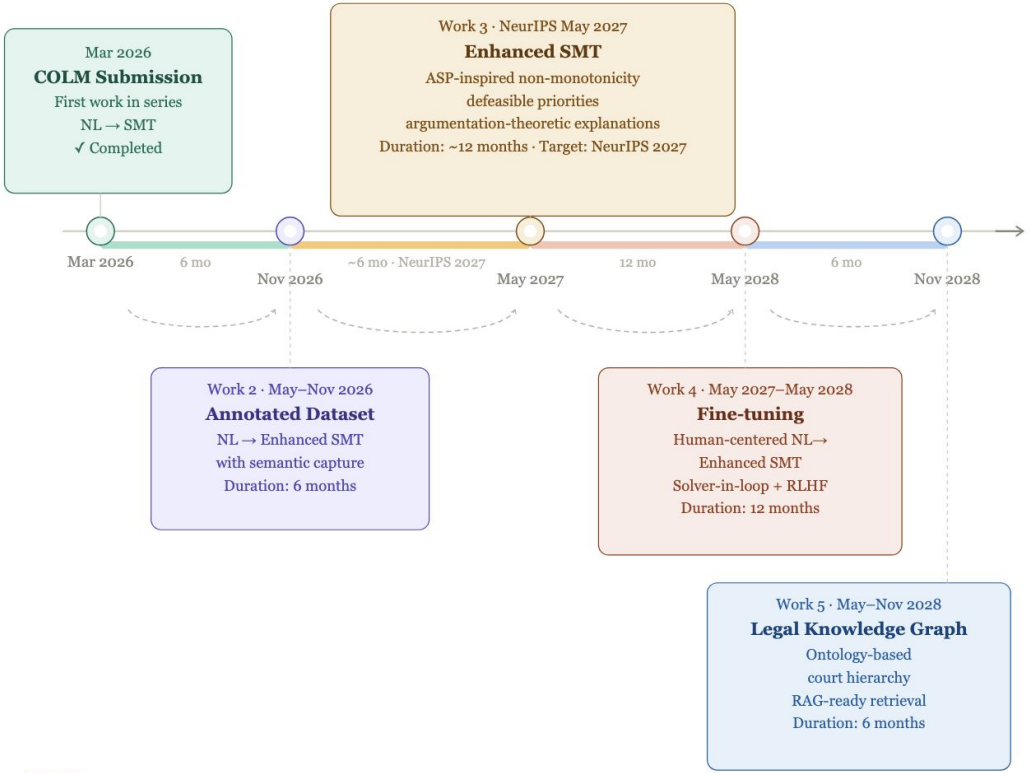
- The goal is to create lawyer-centered assistive Legal AI tools that shape the NextGen legal practice
  - Grounding verification for counsels and judges
  - Neural-symbolic tools that use AI to help with formal verification but leave the legal interpretation to human lawyers
  - Trustworthy, reliable and overall just better legal AI solutions

Q & A

Olivia Wang, pwang95@ucsc.edu

UNIVERSITY OF CALIFORNIA SANTA CRUZ

Future Work and Timeline



Bridging Legal Interpretation and Formal Logic

Olivia Wang, pwang95@ucsc.edu

UNIVERSITY OF CALIFORNIA SANTA CRUZ