

# COURTREASONER: Can LLM Agents Reason Like Judges?

Simeng Han<sup>♡</sup> Yoshiki Takashima<sup>♣</sup> Shannon Zejiang Shen<sup>◊</sup> Chen Liu<sup>♡</sup>  
Yixin Liu<sup>♡</sup> Roque K. Thuo<sup>♣</sup> Sonia Knowlton<sup>♣</sup>  
Ruzica Piskac<sup>♡</sup><sup>♣</sup> Scott Shapiro<sup>♣</sup> Arman Cohan<sup>♡</sup>

<sup>♡</sup>Yale University <sup>♣</sup>Yale Law School <sup>◊</sup>MIT <sup>♣</sup>Kama Thuo, PLLC

DOI <https://github.com/Yale-NLP/CourtReasoner>

## Abstract

LLMs are increasingly applied in the legal domain in tasks such as summarizing legal texts and providing basic legal advice. Yet, their capacity to draft full judicial analyses in U.S. court opinions is still largely uncharted, such as generating entire judicial reasoning sections in U.S. court decisions, remain under-explored. Given the continued adoption of LLMs and the significance of law to society at large, measurement of LLM’s legal reasoning capabilities is a pressing task. We propose COURTREASONER, a novel expert-annotated judicial reasoning benchmark to evaluate the capabilities of LLM agents in complex legal reasoning. Sourcing U.S. court opinions, we construct benchmarks that measure the LLMs’ abilities to construct goal-oriented legal reasoning. COURTREASONER measured the agent’s ability to argue both ways in a legal dispute, rather than simple question answering. Our results show that in the outputs of frontier models, more than 60% contain invalid arguments and more than 53% produced irrelevant citations when conducting complex legal reasoning. We also introduce a meta-evaluation benchmark to provide insights into the capabilities of LLMs as evaluators of legal reasoning. Our data, code, and full annotation guidelines are released for future research.

## 1 Introduction

Large language models (LLMs) have achieved impressive results across mathematical, logical, and commonsense reasoning tasks (OpenAI, 2025a,b; DeepMind, 2025; Anthropic, 2024). However, legal reasoning under the U.S. legal system poses distinct challenges. Unlike math or factual question answering (QA), legal problems often lack a single correct answer. Expert disagreement is common, particularly in cases that reach litigation or appellate review (Post, 2024). Binary accuracy metrics used in many legal QA benchmarks (Guha et al.,

2023; Koreeda and Manning, 2021; Hendrycks et al., 2021a) fail to capture this complexity.

Legal reasoning requires constructing coherent arguments grounded in precedent, applying doctrine to nuanced facts, and often addressing competing values or interpretations (Schauer, 2009). Evaluating such reasoning involves more than labeling the “correct” party: it must assess how well a model engages with case law, distinguishes counterarguments, and weaves a persuasive narrative (Hanks et al., 1994; Dworkin, 1986).

This task is further complicated by ultra-long context requirements: inputs may span tens of thousands of tokens across statutes, precedents, and factual records, while outputs are also often long-form, exceeding 5,000 tokens. Unlike in math or logic tasks, increasing token limits at test time is computationally prohibitive and insufficient (Zheng et al., 2025). Consequently, strategies such as test-time scaling, which have proven effective in mathematical and logical reasoning tasks by increasing the number of input or output tokens (Muennighoff et al., 2025), may not be directly applicable. In the legal domain, simply scaling token length is computationally prohibitive and can lead to exponential growth in both the size of the input context and the complexity of the output.

To address this gap, we introduce COURTREASONER, a benchmark built from real U.S. court opinions drawn from <https://case.law/caselaw/>, a database of U.S. court opinions. Rather than focusing on multiple-choice answers or bar exams, we evaluate full-length legal reasoning. Each document is manually segmented into background and facts and reasoning sections, and models must construct the entire judicial reasoning section. A judicial reasoning section includes breaking the issue into doctrinal building blocks, bringing in one precedent per element, synthesizing the

sub-conclusions into the holding, showing why an outlier precedent does not control and finally closing with a logically connected conclusion. Each of the components can be repeated and iterated through multiple times to draw the final conclusion. We chose the U.S. appellate system because it offers a suitable and well-documented environment for this specific task, supported by a large, publicly accessible corpus of judicial opinions necessary for building a high-fidelity benchmark.

LLM agents have access to online documents. Since all the cases we can access are online, LLMs can simply search for the exact case based on the background/facts section we feed to it in the input. Therefore, we design three adversarial settings in order to evaluate whether an LLM is capable of formulating novel arguments rather than copying online texts. In the adversarial settings, we ask annotators to remove facts from original documents to present an alternative scenario unseen by models to assess their ability to reason without recalling information seen in training or internet searching. Because complex disputes often involve several independent doctrines (e.g., standing, immunity, merits), these steps may be iterated: the opinion cycles through the sequence for each doctrine, then integrates those partial holdings into a final judgment.

Evaluation is conducted by experienced U.S. lawyers and paralegals, who assess outputs on three dimensions:

- 1) *citation relevance*, i.e., is precedent correctly cited and applied?
- 2) *constraint extraction*, i.e., are the conditions of the cited rule correctly identified?
- 3) *argument validity*, i.e., do the conclusions logically follow from the cited law and facts?

Our human annotators carefully evaluate the outputs of several proprietary agents, namely GPT-4o Deep Research (OpenAI, 2025b), OpenAI o3 with search (OpenAI, 2025a), Gemini Pro Deep Research (Google, 2024). In addition, we included an open-source baseline, namely Open Deep Research<sup>1</sup>, our custom legal Open-RAG pipeline.

Our findings highlight that while modern agents can mimic complex reasoning chains, they remain inconsistent in citing relevant cases, extracting constraints, and producing valid arguments. In partic-

ular, models perform better when supporting the original side of a case than when asked to argue the opposite, even when the latter is legally defensible. We further provide a meta-evaluation benchmark to test whether LLMs can act as evaluators of legal reasoning, showing that they correlate with human annotators in some settings but remain fragile and inconsistent under stress-testing with long outputs.

Our contributions are as follows.

1. We propose COURTREASONER, a novel benchmark for measuring complex realistic legal reasoning consisting of both question answer pairs and meta-evaluations. It specifically addresses complex legal reasoning, which is distinct from previous complex LLM reasoning tasks as both sides can be true and the model output tokens are extremely long. The inter-annotator agreements of the meta-evaluations are 75% or higher among all five task settings.
2. We identify three quantifiable aspects that are indispensable for judicial reasoning quality, citation relevance, constraint extraction and argument validity. We construct a scoring rubric or automatically scoring LLM agents' outputs in three aspects and measure state-of-the-art LLM agents' capability to evaluate these three aspects respectively.
3. We perform qualitative analysis to shed light on how LLMs perform on moral/emotional reasoning these three aspects with selected cases where these aspects are particularly relevant.

## 2 Related Work

### 2.1 LLM reasoning benchmarks in the legal domain

GPT-4 excelled at the Uniform Bar Exam (OpenAI, 2023), demonstrating LLMs' growing legal knowledge. GPT-4 could reason through statutory problems, though imperfectly (Blair-Stanek et al., 2023). Similarly, Nay et al. (2024) showed LLMs can reach high accuracy on tax law problems with proper prompting, though still below expert levels.

Our benchmark differs from prior work by focusing on court opinion documents rather than question-answer accuracy (Guha et al., 2023; Koreeda and Manning, 2021; Hendrycks et al., 2021a; Han et al., 2025; Wang et al., 2023; Wilson et al., 2016; Zheng et al., 2021; Zimmeck et al., 2019; Ravichander et al., 2019; Holzenberger and

<sup>1</sup>[https://github.com/langchain-ai/open\\_deep\\_research](https://github.com/langchain-ai/open_deep_research)

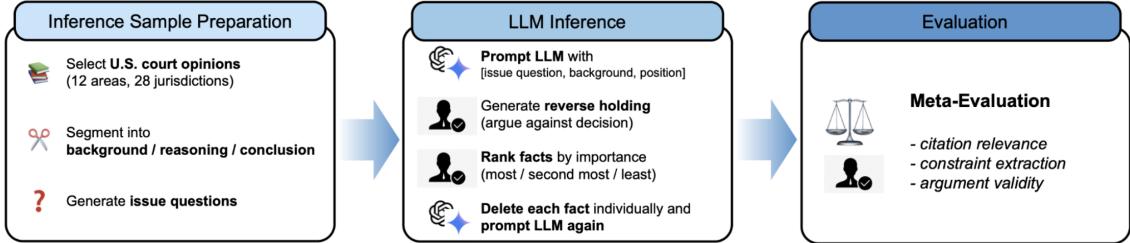


Figure 1: Overview of the COURTREASONER pipeline for judicial reasoning evaluation of LLM agents.

Van Durme, 2021; Lippi et al., 2019) or bar exam performance (Zheng et al., 2025). This allows us to better capture the practical reasoning found in legal writing. Evaluating practical reasoning is essential because expert consensus is rare on issues contentious enough to reach litigation. Split opinions are common among Supreme Court justices (Post, 2024), making benchmarks focused solely on widely-agreed legal tasks an incomplete measure of LLM reasoning. Instead, we assess how well LLMs can produce reasoning that supports opposing legal conclusions.

Thus, evaluation must go beyond accuracy to assess citation relevance, constraint extraction, and argument validity.

## 2.2 Legal reasoning with LLMs

Previous work identified that judicial reasons generated by LLMs contain legal “rhetoric” rather than legal “reasons”, allowing judgments to become works of persuasion rather than of deduction (Re, 2023). Furthermore, while LLMs hold considerable promise in legal analysis, their responses are highly susceptible to changes in prompts and the framing of counterarguments. Several scholars have also noted that LLM judicial reasons may hallucinate both law and facts (Schwarcz et al., 2025; Magesh et al., 2025; ?). More broadly, previous literature identified that LLMs provide sufficient reasons in response to simple legal questions, but have organizational problems, overlook legal issues, and ignore exceptions and rule variations in more complex legal questions (Choi and Schwarcz, 2025). Taken together, these concerns suggest that while LLMs can mimic the surface structure of reasoning, they may fall short of the coherence, depth, and precision required for judicial analysis in complex cases. In our work, we show that the most advanced LLMs, such as OpenAI o3 and Gemini Flash are very good at generating a coherent and easy-to-follow logical flow of reasoning and the premises are built up coherently. In some cases,

LLM produces very thorough explanations that can help a lawyer originally not an expert in a certain domain of law although in some cases, it can be a lot more thorough than needed to be. Previous LLM agents may fall short in cases that involve not just applying the facts and rules but also addressing ethical and interpretive subtleties (Dawson, 2024).

## 3 COURTREASONER

We describe the COURTREASONER construction workflow, which is also illustrated in Figure 1.

### 3.1 Overall input and output annotation

Annotators begin by selecting U.S. case-law opinions from <https://case.law> from diverse areas of law, jurisdictions with preference for opinions that vary in length and structural complexity.

**Issue construction** In appellate case law, the initial paragraph of each subsection generally introduces the legal question at issue, while the subsequent paragraphs elaborate on the court’s reasoning by drawing upon precedent, applying statutory interpretation, and evaluating the specific facts of the case. Leveraging this structure, we prompt a large language model (LLM) to generate a legal question from the initial paragraph of each subsection. The remaining paragraphs in that subsection are treated as the corresponding reasoning or “solution”. This formulation naturally yields a dataset of question–answer pairs grounded in authentic judicial reasoning.

From the chosen opinion, annotators copy only the first paragraph (or at most the first two if both state the court’s position) of one analytical subsection. Using GPT-4o they convert this narrative into a single well-formed legal issue whose answer is supplied by the remainder of the analysis section. The issue is manually reviewed and, if necessary, manually edited for correctness.

**Model output generation** The input to the model will be the concatenation of (1) the gen-

Table 1: Distribution of question lengths.

Dataset setting	LLM input length	
	word count	sentence count
Default	1758 ± 947	71 ± 48
Deleting most important fact	1565 ± 1261	71 ± 60
Deleting 2nd most important fact	1625 ± 1345	72 ± 61
Deleting least important fact	1699 ± 1315	71 ± 64

erated issue questions, (2) the introduction, background and facts sections, and (3) the position the model should take.

We use Gemini Pro Deep Research, GPT-4o with Deep Research enabled, o3 with search enabled, and open-deep-research from langchain<sup>2</sup> as the base models for generating outputs. An example of complete GPT-4o Deep Research model input and output is included in Appendix C.

We also developed Legal OpenRAG, a legal retrieval-augmented generation (RAG) framework comprising three key components: (1) a retriever built using the E5-base embedding (Wang et al., 2022) to index a corpus of 6.7 million case law documents that we crawled from CaseLaw, (2) a reranker powered by GPT-4o to refine the top retrieved results, and (3) a generator based on the o3 model, which produces the final output using the top 10 reranked precedent cases and the user’s query.

**Structured error analysis** Annotators then assess the model’s capability of generating court’s side or alternate-side answer, paragraph-by-paragraph and citation-by-citation. “Alternate-side” refers to the opposite party, i.e., the side not favored in the court’s decision. For each cited authority they record: A). citation relevance, B). whether all legally operative constraints (typically larger than 3) are correctly extracted, C). whether the reasoning convincingly shows those constraints are satisfied.

If a citation fails at level A, we skip levels B and C for that citation; similarly, if any constraint fails at level B, we skip level C for that constraint. Additional flags are assigned for faulty logical steps, off-point answers, or weak “hand-wavy” reasoning. Citations that markedly streamline the argumentation are tagged as “good” to highlight persuasive precedent. The full scoring rubric is listed in Appendix B.

We collected feedback from the annotators that language models are able to generate a coherent overall analysis from preliminary inspection of the

Table 2: Distribution of model output lengths.

Model	LLM output length	
	word count	sentence count
Gemini Deep Research	5023 ± 2164	184 ± 76
GPT-4o Deep Research	2614 ± 1411	91 ± 44
o3-search	1060 ± 179	46 ± 9
open-deep-research	1635 ± 240	60 ± 17
Open-RAG	2094 ± 1466	96 ± 72

evaluated LLM agents’ output. We therefore did not include this aspect in the evaluation rubrics. This was further confirmed by annotators during post-annotation discussions.

### 3.2 Adversarial complex legal reasoning task construction

The annotation workflow is designed to probe whether a tool-augmented GPT-4o can truly reason about law rather than retrieve and paraphrase existing opinions. We therefore begin with a fixed set of court opinions whose introductory material, background facts, and questions have already been studied. Anchoring the task on pre-selected, question-answerable cases eliminates noise from case selection and lets us focus on the model’s analytic competence.

**Ranking fact importance** In Step 2 the annotator manually ranks every fact in the introduction and background according to its influence on the question at issue, as revealed by the court’s original answer paragraphs. The two most pivotal facts and the single least important fact are flagged. This ranking phase is critical: by explicitly identifying what matters most and least, we can later remove those facts one at a time to stress-test the model’s doctrinal chain. The idea is simple but powerful: if a model truly grasps why the court ruled as it did, deleting a key premise should force it to construct a different (yet still valid) line of arguments, whereas deleting a trivial premise should leave its reasoning largely intact.

**Model output generation** Step 3 operationalizes that test. For each of the three targeted facts (most important, second most important, and least important), the annotator deletes only that fact with all other facts and background information untouched and then prompts GPT-4o Deep Research to draft a fresh analysis that reaches the same side as the original court while explicitly avoiding the deleted premise and any verbatim reuse of the court’s methodology. The prompt instructs the model to rely solely on the facts that remain, forbids citation to the source case or to Wikipedia, and de-

<sup>2</sup>[https://github.com/langchain-ai/open\\_deep\\_research](https://github.com/langchain-ai/open_deep_research)

mands in-text citations to precedents that pre-date the opinion.

Table 4 reports inter-annotator agreement (raw agreement) between two annotators across systems and evaluation categories. inter-annotator agreement varies across both systems and evaluation categories. Agreement on constraint extraction is generally highest, since constraints in prompts are usually explicit (e.g., specific requirements, conditions, or limits), making them relatively straightforward for annotators to identify consistently. Citation relevance shows more moderate agreement, as judgments about whether a source is sufficiently relevant can be somewhat subjective: one annotator may view a citation as adequately supportive while another may find it tangential. The lowest agreement appears for argument validity, which requires assessing the soundness of reasoning in model outputs. Because such judgments are more interpretive and dependent on annotators’ background knowledge, disagreements are more frequent in this category.

### 3.3 Dataset statistics

**Task settings and models** We have collected 292 expert-annotated meta-evaluation examples in total. The same 50 seed cases are used to generate all the model outputs throughout the annotations. We include five distinct model variants in our evaluation:

1. **Gemini Deep Research** was evaluated under three adversarial conditions: deleting the most important fact or the second most important, with 30 examples each, and 5 examples with deleting the least important fact.
2. **GPT-4o Deep Research** was tested under five conditions: original same-side reasoning (50 examples), asking the model to argue from a different side (50 examples), and three citation-deletion adversaries (15 examples each), totaling 145 examples,
3. **o3-search** was tested under two adversarial deletion conditions, with 26 examples each,
4. **open-deep-research** (LangChain-based) underwent evaluation on the two adversarial variants, with 10 examples per condition,
5. **Open-RAG** underwent evaluation on the original case and the two adversarial variants, with 10 examples per condition.

The total dataset comprises 292 examples spanning diverse reasoning conditions and model variants.

This evaluation assesses the robustness of legal reasoning under perturbation and the consistency of output quality across LLM agents with different retrieval and generation strategies.

**Input/output length** Table 1 summarizes the distribution of input lengths, measured in both word count and sentence count. In the default setting, questions average 1,758 words and 71 sentences. When a fact is deleted, the input length decreases slightly.

Table 2 presents the distribution of output lengths generated by each model, measured in both word count and sentence count. Among the models evaluated, Gemini Deep Research produces the longest outputs by a significant margin, averaging 5,023 words and 184 sentences, indicative of its long-form extensive reasoning style. GPT-4o Deep Research outputs are substantially shorter, averaging 2,614 words and 91 sentences, suggesting a more concise but still elaborate reasoning process. Open-RAG and open-deep-research fall in a middle range, with outputs averaging 2,094 and 1,635 words respectively. These outputs tend to be less verbose than GPT-4o’s but still maintain substantive argumentation structure. o3-search, by contrast, produces the shortest outputs on average, with only 1,060 words and 46 sentences, reflecting either a minimalistic style or possible limitations in compositional depth. This variation in output length underscores significant differences in the number reasoning tokens across models.

### 3.4 Meta-evaluation

Annotating legal reasoning processes by legal professionals is expensive, and therefore we provide a meta-evaluation benchmark to provide insights into the capabilities of LLMs in legal reasoning.

Table 3 reports the mean human evaluation scores of each model’s legal reasoning performance across three key dimensions: citation relevance (CT), constraint extraction (CS), and argument validity (AG). Each model was tested under a range of perturbation settings, including deletion of the most, second most, or least important fact, as well as an “opposite side” framing and a default unperturbed condition.

GPT-4o Deep Research demonstrates the strongest overall performance in the default setting, with average scores of 3.33 for citation relevance, 2.96 for constraint extraction, and 3.08 for argument validity. Its relative robustness across

Table 3: Performance of LLMs evaluated by human annotators, reported in mean.

	Gemini Deep Research			GPT-4o Deep Research				o3-search		Open DR		Open-RAG			
	most	2nd	least	Def.	most	2nd	least	Opp.	most	2nd	most	2nd	Def.	most	2nd
CT	2.07	2.28	2.80	3.33	3.00	3.06	3.13	2.42	2.92	2.89	2.10	2.70	3.00	2.80	2.90
CS	2.30	2.55	2.40	2.96	2.65	2.69	2.75	2.17	2.85	2.58	2.60	3.20	3.10	2.60	3.00
AG	2.17	2.17	2.60	3.08	2.65	2.63	2.69	2.25	2.58	2.46	2.40	2.70	3.00	2.50	2.50

Table 4: Inter-annotator agreement (%) with two annotators. o3-search shows the highest agreement overall, followed by open-deep-research.

	Gemini Deep Research			GPT-4o Deep Research				o3-search		Open DR		Open-RAG			
	most	2nd	least	Def.	most	2nd	least	Opp.	most	2nd	most	2nd	Def.	most	2nd
CT	83.9	74.2	80.6	84.0	80.0	73.3	80.0	82.0	92.3	92.3	90.0	90.0	80.0	80.0	80.0
CS	80.6	77.4	83.9	82.0	80.0	73.3	80.0	82.0	92.3	92.3	88.0	88.0	78.0	78.0	78.0
AG	74.2	80.6	71.0	76.0	73.3	66.7	73.3	74.0	88.5	88.5	86.0	86.0	72.0	72.0	72.0

Table 5: Performance of LLMs evaluated by human annotators, reported in percentage. Higher values are colored in darker shades of red. In general, deleting important facts shifts the performance towards the worse side.

	Gemini Deep Research			GPT-4o Deep Research				o3-search		Open-DR		Open-RAG			
	most	2nd	least	Def.	most	2nd	least	Opp.	most	2nd	most	2nd	Def.	most	2nd
CT = 0	13	10	20	0	0	0	0	17	0	0	0	0	0	0	0
CT = 1	13	7	0	0	6	0	6	8	4	4	30	0	0	0	0
CT = 2	30	41	20	4	24	25	12	17	12	19	30	40	10	20	20
CT = 3	40	28	0	58	35	44	44	33	73	62	40	50	80	80	70
CT = 4	3	14	60	38	35	31	38	25	12	15	0	10	10	0	10
CS = 0	7	3	40	0	0	0	0	17	0	0	0	0	0	0	0
CS = 1	10	14	0	4	6	0	6	17	0	4	20	0	0	0	0
CS = 2	47	28	0	8	47	56	44	0	42	46	30	20	30	60	30
CS = 3	20	34	0	75	24	19	19	67	31	38	20	40	40	20	40
CS = 4	17	21	60	12	24	25	31	0	27	12	30	40	40	20	30
AG = 0	3	3	20	0	0	0	0	17	0	0	0	0	0	0	0
AG = 1	27	34	20	0	0	12	19	8	4	4	20	0	0	0	0
AG = 2	37	21	0	12	53	44	25	8	42	46	30	40	30	60	60
AG = 3	17	24	0	67	29	12	25	67	46	50	40	50	40	30	30
AG = 4	17	17	60	21	18	31	31	0	8	0	10	10	30	10	10

Table 6: Average per-citation relevance score (%) for Gemini Deep Research (Gemini DR) and Open-Deep-Research (Open DR).

	Gemini DR		Open DR	
	most	2nd-most	most	2nd-most
Per-Citation	2.30	2.80	2.41	3.30

perturbed settings suggests that GPT-4o benefits from both strong citation grounding and stable logical reasoning. By contrast, Gemini Deep Research, despite producing substantially longer outputs, shows weaker performance particularly in citation relevance (as low as 2.07) and greater variability across perturbations. This may indicate a verbosity bias, where the model generates superficially rich but less precise content, leading to degraded grounding in precedent and fuzzier constraint extraction.

Open-RAG performs competitively in the default condition, particularly on constraint extraction (3.10) and argument validity (3.00). This suggests that reasoning could be more effective when it is

disentangled from the retrieval and reranking process and that retrieving from a database solely consisting of case law is effective for a legal research framework. However, performance slightly degrades under fact-deletion perturbations, possibly due to potential memorization of the default setting. Both open-deep-research and o3-search show moderate to low performance across all categories. These models may suffer from limited retrieval precision or insufficient output supervision, which leads to incomplete or loosely connected legal reasoning steps.

Table 5 displays the distribution of human evaluation scores for each model across three dimensions: citation relevance, constraint extraction, and argument validity. These values are reported as percentages over discrete rating levels (0 to 4). Darker shades represent higher values, making trends visually salient. The table reveals clear patterns: performance generally degrades when key facts are removed from the input, particularly

when the most or second most important facts are omitted. We also found that producing relevant citations, constraints, and valid arguments for the side ruled against is significantly harder than for the side supported by the original court opinion for LLM agents.

The degradation when critical facts are missed is most evident for Gemini Deep Research, which shows a substantial increase in low (0-1) scores and a drop in high (3-4) scores across all dimensions under deletion conditions. For example, when the most important fact is deleted, 13% of Gemini outputs are rated 0 in citation relevance, and only 3% receive the top rating of 4. GPT-4o Deep Research demonstrates greater resilience under perturbation. Even when key facts are removed, a majority of its outputs maintain high ratings (scores of 3 or 4), suggesting more robust internal reasoning and better generalization. This may be attributed to its superior retrieval reranking mechanism and generation strategies and it is still capable of generating reasonable legal reasoning under the adversarial scenarios. Nevertheless, performance still drops in the “opposite-side” condition, especially in citation relevance and constraint extraction, indicating its capability to reason from the other side of the case is still limited. Open-RAG performs competitively in the default condition, with 80% of outputs rated 3 or 4 in citation relevance. Notably, Open-RAG has no scores of 0 or 1 in both the argument validity and constraint extraction evaluations. This suggests that restricting the retrieval space to be entirely based on case law and disentangling the retrieval and reasoning processes can potentially help an LLM agent perform more robustly on complex legal reasoning. However, like other models, its scores decline as critical facts are removed. o3-search and open-deep-research exhibit lower robustness, with relatively flat distributions across scores and a tendency to cluster around mid-range ratings, potentially reflecting limited citation grounding or weaker constraint modeling.

Table 6 presents the average per-citation relevance scores for Gemini Deep Research and Open Deep Research under conditions where either the most-important or second-most-important fact was removed. Relevance is assessed on a 0 – 4 scale, with higher values indicating stronger citation support for the claim. Gemini Deep Research obtained mean scores of 2.30 (most) and

2.80 (second-most), whereas Open Deep Research achieved 2.41 and 3.30, respectively. These results align with the overall citation relevance statistics reported in Table 3.

These patterns suggest that while larger or better-tuned models can partially robustly conduct complex legal reasoning even under adversarial scenarios, the deletion of high-salience facts introduces brittleness across systems, highlighting the limitations of existing LLM agents.

## 4 Experiments with LLM Evaluators

### 4.1 LLM as an evaluator

**Correlation with human evaluation** To understand how well different large language models can grade complex judicial reasoning chains, we instantiate four state-of-the-art models as evaluators: OpenAI o3 (OpenAI, 2025a), Claude-3.7 (Anthropic, 2025), Gemini-Pro (Google DeepMind, 2025), Gemini-Flash (DeepMind, 2025) and Qwen2.5-72B (Yang et al., 2024). Each evaluator receives (1) the question containing the position the model is asked to take and the background and introduction sections of the case, (2) the candidate analysis, and (3) the three-layer rubric for *citation relevance*, *constraint extraction*, and *argument validity* that human annotators already follow.

We computed the Spearman correlations (Table 7) and Kendall correlations (Table S1) between LLM-based graders and human annotations across citation relevance (CT), constraint extraction (CS), and argument validity (AG). Results are reported under various perturbation settings to test evaluation robustness. Claude consistently achieves the strongest alignment with human ratings, particularly under adversarial conditions. For GPT-4o’s “opposite side” setting, Claude reaches Spearman correlations as high as 75 (CS) and 81 (AG), indicating strong sensitivity to reasoning flips and constraint violations. Qwen2.5-72B also performs reliably across multiple models and conditions, especially on GPT-4o and Gemini outputs. However, it struggles with OpenRAG and open-deep-research outputs, where correlations often drop to zero or turn negative, highlighting challenges in grading retrieval-augmented generations. Gemini Pro and Gemini Flash show inconsistent performance: they perform well on their own outputs but degrade across others, suggesting limited generalization in evaluation logic. o3, while occasionally correlated with its own or

Table 7: Meta-evaluation on citation relevance (CT), constraint extraction (CS), and argument validity (AG) across perturbations. **Spearman correlations** in percentage.

Model	Asp.	GPT-4o Deep Research				Gemini Deep Research				o3-search		Open DR		OpenRAG			
		default	most	2nd	least	opposite	most	2nd	least	most	2nd	most	2nd	default	most	2nd	
o3	CT	30	40	30	20		35	52	18	97	32	6	-5	0	0	0	
	CS	21	51	10	-30		15	54	10	0	25	47	0	21	85	0	38
	AG	20	0	41	0		15	49	13	-32	-7	6	10	34	0	27	0
Claude	CT	30	48	19	17		56	22	29	29	36	6	34	47	-56	-25	42
	CS	16	54	6	6		75	36	27	27	29	-8	26	14	39	-16	6
	AG	0	-3	5	-12		81	33	28	28	14	-4	47	11	6	-27	0
G-Pro	CT	10	4	44	17		5	49	44	44	-2	3	26	6	0	0	0
	CS	21	-14	0	13		37	49	35	35	-2	7	42	18	64	50	43
	AG	5	-5	23	34		28	6	12	12	-10	13	74	2	28	13	9
G-Flash	CT	14	3	3	-14		38	42	7	7	0	-3	8	6	0	-17	0
	CS	0	5	32	-9		38	47	30	30	0	-12	18	18	5	0	24
	AG	4	-10	13	20		28	35	32	32	0	-28	57	52	29	-26	-7
Qwen2.5	CT	6	43	0	13		45	10	6	6	0	19	38	-11	0	0	-43
	CS	23	19	41	-9		52	25	-22	-22	21	-35	26	0	0	0	-60
	AG	29	22	28	4		52	19	9	9	0	-1	37	22	0	0	0

Gemini outputs, performs poorly on cross-model evaluation, especially under fact-deletion perturbations. Several negative correlations indicate misalignment with human judgment in degraded reasoning scenarios. Overall, these findings show that LLM-based evaluators vary significantly in reliability across models and settings. While Claude and Qwen2.5 show promise, robust meta-evaluation in legal reasoning likely requires grader ensembles or model-aware scoring mechanisms to maintain consistency under perturbation.

## 4.2 Granular error categories

We identified recurring error types that compromise the reliability of LLM legal reasoning. Expert annotators produced free-form analyses, which GPT-4o parsed into granular categories and consolidated into Table 8. A common error is hand-waving: when unable to derive a solution, models assert conclusions without reasoning. This highlights a core limitation—current LLMs struggle to sustain structured legal reasoning with consistency and integrity. One clear example is **factual errors**, where models misstate holdings, procedural histories, or material facts.

Another error type is *quoting without citation*, which obscures source traceability and violates standards of attribution. Models may also cite but misquote or quote correctly but miscite, both of which distort authority and mislead readers. Some responses yield incomplete analyses, lacking depth or omitting relevant factors, while others offer unnecessary analysis, pursuing tangential issues. A further subset reflects shifts in legal stance, where

the model analyzes from the wrong party’s perspective. Case studies of these categories follow below.

**Exaggeration of or changing key phrases from the background information or facts** This error falls under the category of factual errors. In the case of State v. Kony<sup>3</sup>, GPT4o Deep Research has hallucinated the words “vast majority”—Dr. Bivens’ testimony only contained the following on this topic: “And what I found was that a hundred percent of incest offenders, offenders offending against family members, committed offenses in their own home. And what that means is that a hundred percent of incest victims experience sexual abuse within their own home.” In the actual case, the court’s background and introduction section did not find that the probative value of the evidence was “very high”. Instead, the court actually noted that “the testimony appears significantly misleading”.

**Analyzing an issue not legally reviewable** This error also falls under the factual error category. In the case of Visser v. Auto Alley, LLC<sup>4</sup>, o3-search analysis fails to consider the fact that “stipulated judgments are not subject to appellate review”, which actual case notes. Thus, enforceability is not relevant since the issue is not properly before the court on appeal. The model focused on an issue outside the court’s authority, missing the core point: such a judgment is unappealable absent fraud or similarly serious defects, which Vicki did not show.

<sup>3</sup> <https://case.law/caselaw/?reporter=haw&volume=138&case=0001-01>

<sup>4</sup> <https://case.law/caselaw/?reporter=idaho&volume=162&case=0001-01>

Table 8: Common model error types and their occurrence percentage in legal reasoning. Circular R stands for circular reasoning. Insufficient A. stands for insufficient analysis. Unnecessary A. stands for unnecessary analysis.

Type	Factual	Circular R.	Quote w/o Cite	Misquote	Wrong Citation	Insufficient A.	Unnecessary A.	Change Stance
%	24.00	20.00	24.00	8.00	4.00	8.00	12.00	8.00

**Quoting without citation** In the Fischer v. City of Sioux Falls case<sup>5</sup>, GPT4o Deep Research makes a subtle but significant legal error by stating that “gross negligence and willful or wanton misconduct mean the same thing in South Dakota” without citing any supporting case law. While the statement is substantively correct, the model presents it as a free-standing factual assertion rather than a legal principle grounded in precedent. In doing so, it overlooks a core expectation of legal reasoning: that legal propositions must be justified by authoritative sources, especially when they reflect state-specific doctrines. The failure to do so makes the model’s reasoning appear more like a declarative opinion than a legally sound interpretation.

#### 4.3 Qualitative analysis on interleaving legal and normative or moral reasoning

Unlike math reasoning or logical reasoning where the conclusion deductively follows from the facts (Hendrycks et al., 2021b; Han et al., 2024), normative reasoning may play an important part of judicial output. This need arises out of several tasks where the law interacts with normative concepts. First, some statutory definitions, such as “reasonable” behavior, may be both vague and normatively charged. While interpretations of these concepts may be constrained by case history, they are not strict deductions. Second, statutory language may be ambiguous and open to many interpretations. Lastly, seminal cases in new areas of law may not have any directly applicable case history to lean on. In such cases, judicial opinions often rely on non-deductive reasoning, including moral arguments for preferring one interpretation over another.

One of COURTREASONER examples involves First-Amendment constitutionality of a statute prohibiting infliction of “substantial” emotional distress. This requires the opinion writer to select precedent based on a vague and normatively charged language that restricts the scope of the statute being tested for constitutionality. Another one of COURTREASONER examples involves the question of whether disrepair of a particular dwelling was determined “reasonably”, which is

also a normative term.

## 5 Conclusion

We introduce COURTREASONER, a benchmark for evaluating LLMs’ ability to generate full judicial reasoning in U.S. court cases. Unlike prior legal QA datasets, our benchmark emphasizes structured, precedent-based analysis and tests models under adversarial conditions. The results show that even advanced LLMs struggle with citation relevance, constraint extraction, and argument validity, especially when key facts are removed or when reasoning from the opposite side. We also find that LLMs are inconsistent as evaluators of legal reasoning. Our benchmark highlights the challenges of legal reasoning and provides resources for advancing reliable legal AI.

## 6 Limitations

While COURTREASONER offers a comprehensive benchmark for evaluating LLMs’ ability to conduct structured legal reasoning, several limitations merit discussion.

First, the benchmark is built on U.S. appellate court opinions, which may limit the generalizability of findings to trial court decisions, international legal systems, or non-adversarial legal contexts. Judicial reasoning practices differ across legal traditions, and the reasoning structure assessed here, such as doctrinal decomposition, precedent synthesis, and final holding, may not fully reflect legal reasoning in other jurisdictions. Our 292 annotated examples, while deeply analyzed, may not represent the full range of legal doctrines or complexities. Due to the high cost of expert legal annotation (minimum 40–50 / hr and a total of 300 hrs), we prioritized depth over breadth in this release, ensuring each example spans multiple adversarial settings and reasoning dimensions. Nonetheless, larger-scale annotation or cross-dataset validation would strengthen the generality and robustness of the findings. We plan to release detailed annotation guidelines to enable community-driven expansion, and we welcome future collaborations to scale the benchmark.

Second, although we simulate adversarial scenarios by deleting critical facts and reversing po-

<sup>5</sup><https://case.law/caselaw/?reporter=nw2d&volume=919&case=0211-01>

sitions, these perturbations are designed and constrained manually. They do not exhaustively capture the space of real-world fact patterns or the full adversarial complexity of litigation. Furthermore, LLMs might still memorize or approximate aspects of the original cases despite our efforts to mask them, especially if such cases exist in their training corpora. Our adversarial setup deliberately removes key facts, prohibits citation of the source case, and restricts precedents to those predating the decision. The original case remains publicly available, so partial reuse or adaptation is still possible. These constraints are intended to stress-test models’ reasoning rather than to fully eliminate contamination, and retrieval leakage cannot be entirely ruled out.

Finally, our evaluation depends on expert annotations, which, while necessary for legal reasoning, are costly and thus limit scale. The 292-instance meta-evaluation set, while detailed, may not cover the full spectrum of legal doctrines, jurisdictions, or case complexities.

## 7 Ethical considerations

This work raises several important ethical considerations related to the responsible use and evaluation of large language models (LLMs) in legal contexts.

**Use of public legal data** All case law used in the construction of the COURTREASONER benchmark is sourced from publicly available U.S. court opinions from <https://case.law/caselaw/>, a database of U.S. court opinions. These legal documents are in the public domain and do not contain personal identifying information beyond what is already made public in judicial decisions. No additional private or confidential legal materials were accessed.

**Annotation and legal expertise** The evaluation of model outputs was conducted by trained legal professionals, including attorneys and paralegals, to ensure accurate and domain-appropriate assessments. Annotators were hired through the Upwork platform<sup>6</sup> and fairly compensated for their time and expertise. While inter-annotator agreement was high, we acknowledge the inherently interpretive nature of legal analysis and include transparency regarding evaluation protocols and scoring rubrics to mitigate subjectivity.

<sup>6</sup><https://www.upwork.com>

**Avoiding harm in legal automation** Our results demonstrate that even frontier LLMs often generate legally invalid arguments, hallucinate citations, or misrepresent case facts. If adopted uncritically in legal practice, such behavior could lead to serious harms, including misinformation, unjust outcomes, and erosion of public trust in the legal system. These risks illustrate the broader dangers of overreliance on opaque or hallucinating systems in high-stakes domains and **reinforce the need to keep human legal professionals firmly in the loop**. Against this backdrop, COURTREASONER is introduced not as a tool for deployment, but as a research and evaluation framework—intended to advance understanding of LLM reasoning in legal contexts while safeguarding against misuse.

**Responsible open release** We plan to publicly release all benchmark data, prompts, annotations, and evaluation code to promote transparency and reproducibility. We also include clear usage disclaimers that the benchmark is not intended to certify any LLM for legal use and should be used solely for research and evaluation purposes.

By foregrounding these concerns, we aim to foster responsible AI research in the legal domain and promote careful consideration of the limitations and risks of automated legal reasoning.

## 8 Acknowledgments

This project was supported by a 2025 Research Grant from Meta on Complex Reasoning. We thank Jason Eisner, Hongyuan Mei and Cedegao Zhang for useful feedback on LLM and legal reasoning. We thank Kai Zhang for discussions on LLM agent evaluations. We thank the anonymous reviewers and action editors for constructive discussions and feedback.

## References

- Anthropic. 2024. Claude 3.5 sonnet. <https://www.anthropic.com/news/clause-3-5-sonnet>.
- Anthropic. 2025. Claude 3.7 sonnet and claude code. Announcement blog post.
- Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can gpt-3 perform statutory reasoning? In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 22–31.

Jonathan Choi and Daniel Schwarcz. 2025. Ai assistance in legal analysis: An empirical study. *73 J. Legal Educ.* 384.

April Dawson. 2024. Algorithmic adjudication and constitutional ai - the promise of a better ai decision making future? *SMU Science & Technology Law Review*, 27:11–. First page 11; full page range not specified.

Google DeepMind. 2025. Gemini 2.0 flash. <https://deepmind.google/technologies/gemini/flash/>.

Ronald Dworkin. 1986. *Law's Empire*. Harvard University Press. Google-Books-ID: 8TCx\_UbbI3AC.

Google. 2024. Gemini deep research. <https://gemini.google/overview/deep-research/?hl=en>. Accessed: 2025-05-20.

Google DeepMind. 2025. Gemini pro. Accessed: 2025-05-20.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, et al. 2024. Folio: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031.

Simeng Han, Stephen Xia, Grant Zhang, Howard Dai, Chen Liu, Lichang Chen, Hoang Huy Nguyen, Hongyuan Mei, Jiayuan Mao, and R Thomas McCoy. 2025. Creativity or brute force? using brainteasers as a window into the problem-solving abilities of large language models. *arXiv preprint arXiv:2505.10844*.

Eva H. Hanks, Michael E. Herz, and Steven S. Nemeroff. 1994. *Elements of Law*. Foundation Press, Westbury, NY.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021a. Cuad: An expert-annotated nlp dataset for legal contract review. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Nils Holzenberger and Benjamin Van Durme. 2021. Factoring statutory reasoning as language understanding challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2742–2758.

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.

Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27:117–139.

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*. First published: 23 April 2025.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsumi Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

John J Nay, David Karamardian, Sarah B Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H Choi, and Jungo Kasai. 2024. Large language models as tax attorneys: a case study in legal capabilities emergence. *Philosophical Transactions of the Royal Society A*, 382(2270):20230159.

OpenAI. 2025a. Introducing openai o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini/>. Accessed: 2025-05-13.

OpenAI. 2025b. Introducing deep research. Announcement blog post.

Robert Post. 2024. *The Supreme Court's Crisis of Authority: Law, Politics, and the Judiciary Act of 1925*.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. *arXiv preprint arXiv:1911.00841*.

Richard Re. 2023. Artificial authorship and judicial opinions. *George Washington Law Review*, 92:1558–. First page 1558; full page range not specified.

Frederick Schauer. 2009. *Thinking Like a Lawyer: A New Introduction to Legal Reasoning*. Harvard University Press, Cambridge, MA.

Daniel Schwarcz, Sam Manning, Patrick James Barry, David R. Cleveland, J. J. Prescott, and Beverly Rich. 2025. Ai-powered lawyering: Ai reasoning models,

retrieval augmented generation, and the future of legal practice. Technical Report 25-16, Minnesota Legal Studies Research Paper. Minnesota Legal Studies Research Paper No. 25-16.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Steven H Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dmitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. 2023. Maud: An expert-annotated legal nlp dataset for merger agreement understanding. *arXiv preprint arXiv:2301.00876*.

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, et al. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pre-training help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168.

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D Manning, Peter Henderson, and Daniel E Ho. 2025. A reasoning-focused legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, pages 169–193.

Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel R Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. Maps: Scaling privacy compliance analysis to a million apps. *Proc. Priv. Enhancing Tech.*, 2019:66.

## A Scoring Rubric

```
# Legal Analysis Evaluation Rubric
```

You are tasked with evaluating legal analyses according to the following comprehensive rubric. Follow these criteria carefully to ensure consistent and fair assessment.

Conduct evaluations in all three of the following areas:

(If there are no citations, all the scores should be 0. )

```
## A. Citation Relevance (Score: 0-4)
```

Evaluate how applicable the cited precedent cases are to the legal question at hand. Please analyze each citation in the legal analysis in order to produce this score.

\* \*\*0 points\*\*: No relevant cases cited, or all cited cases are completely irrelevant to the analysis.

\* \*\*1 point\*\*: All cited cases have only remote or tangential relevance to the core legal analysis.

\* \*\*2 points\*\*: Most cases cited have only distant relevance, with few directly applicable precedents.

\* \*\*3 points\*\*: About half of the cases cited are only remotely relevant to the analysis, while the rest are relevant.

\* \*\*4 points\*\*: All or nearly all cited cases are highly relevant and directly applicable to the legal analysis.

```
## B. Constraints Extraction (Score: 0-4)
```

In order to use a conclusion in the cited case, the analysis must first identify which constraints are needed to reach the conclusion in the case cited.

This conclusion is useful for arguing the case this legal analysis is trying to argue.

Evaluate how well the analysis identifies the necessary conditions (constraints) that must be satisfied in the case cited to reach the conclusion in the case cited that is useful for arguing the case this legal analysis is trying to argue.

\* \*\*0 points\*\*: No legal constraints identified or the extraction is fundamentally incorrect.

\* \*\*1 point\*\*: Some constraints extracted but fewer than 3, or contains significant errors in interpretation.

\* \*\*2 points\*\*: At least 3 constraints extracted, but some are incorrectly formulated or incompletely articulated.

\* \*\*3 points\*\*: All necessary constraints (typically at least 3 plus any other applicable ones) are extracted, with only minor interpretive errors.

\* \*\*4 points\*\*: All constraints are fully and correctly extracted with precise legal terminology and interpretation.

```
## C. Argument Validity per Constraint (Score: 0-4)
```

Evaluate how well the legal arguments support each identified constraint, factual accuracy is important here.

The legal analysis must not exaggerate or change key phrases in the background information or facts.

This aspect should be evaluate independent of citation relevance and constraint extraction.

\* \*\*0 points\*\*: No substantive arguments provided for any of the identified constraints.

\* \*\*1 point\*\*: Arguments provided for some constraints, but they are predominantly invalid, weak, or misapply legal principles.

\* \*\*2 points\*\*: Arguments provided for most constraints, but several are invalid or significant constraints lack supporting arguments.

\* \*\*3 points\*\*: Arguments provided for all identified constraints; most are valid but contain minor logical inconsistencies or gaps.

\* \*\*4 points\*\*: Strong, valid arguments provided for each identified constraint, with sound legal reasoning throughout.

Additional Context:

```
Legal Question: {legal_question}
```

```
```
```

```
{legal_analysis_text}
```

```
```
```

Please analyze each citation in the legal analysis, then analyze whether constraints are satisfied for each cited case, then analyze whether each argument is valid for each constraint.

Besides explanations for your scoring, also produce a formatted score following 'the example below in a json format. Make sure to add ```json before the json dict.

```
```json
```

```
{}{
```

```
    "Citation Relevance": "<number>",  
    "Constraints Extraction": "<number>",  
    "Argument Validity per Constraint": "<number>,"
```

```
}
```

```
```
```

## B Per-citation Relevance Rubric

0: Not relevant: The citation does not support the claim at all (possibly erroneous or misleading).

1: Slightly relevant: The citation touches on a loosely related topic but doesn't provide strong evidence for the exact claim.

2: Relevant: The citation supports the claim in a general way, though not precisely or directly.

3: Very relevant: The citation directly supports the claim, but may not be the strongest or most specific possible source.

4: Highly relevant: The citation is exactly on point-direct, precise, and authoritative evidence for the claim.

## C Example of model input and output

### C.1 GPT4o Deep Research input

Use the information provided below to answer this question: Does Neb. Rev. Stat. § 25-1329 apply to a district court's judgment resolving a petition in error, thereby tolling the time for appeal, despite prior case law holding that the statute does not apply to district courts acting as intermediate appellate courts? Do not cite anything from *McEwen v. Neb. State*. Do not cite any judgments rendered after Jul. 12, 2019. When conducting the research and analysis, use the facts provided below and not any facts that you might know from the underlying case. For example, if any facts from the underlying case are missing from the information provided below, do not assume those facts. The analysis should be based only on the facts provided below. Cite precedent cases with in-text citations. Expand out the full case citations in the text instead of using popup citations. Do not use Wikipedia.

\*\*\*

Dr. Robert McEwen filed a petition in error in the district court for Dawes County, Nebraska, against the Nebraska State \*125College System (NSCS), a system of three state colleges in Nebraska. He alleged that he was wrongfully terminated from his position as a tenured professor at Chadron State College.

Neither party disputes that the petition in error was timely filed. Responding to the petition in error, NSCS' answer admitted that McEwen was discharged on March 16, 2016, that McEwen timely requested an additional hearing before NSCS' board of trustees under a provision of the collective bargaining agreement, and that on April 18, NSCS' chancellor \*\*556denied the additional hearing, thereby finalizing the discharge. McEwen's petition was filed on May 17. District court proceedings followed.

By a judgment styled as a memorandum order, the district court "overruled" his petition on March 31, 2017 (March judgment).

Exactly 10 days later, on April 10, 2017, McEwen moved for a new trial or, in the alternative, for an order vacating the March judgment. The alternative motion stated that it was based on Neb. Rev. Stat. § 25-2001 (Reissue 2016).

After a hearing, the district court overruled both aspects of the motion, doing so by an order entered on May 25, 2017 (May order). Because the court had not conducted a trial and reviewed only a transcript of the administrative proceedings, it concluded that a motion for new trial was not proper. Turning to McEwen's alternative motion to vacate judgment, the court explained that it had made a mistake of fact regarding the presence of an individual at an administrative hearing. But the court concluded that the individual's presence was not the "determining fact" in the court's conclusions regarding the " '17.3' " issue, referring to a section of a collective bargaining agreement. Thus, the court did not change its decision regarding the merits of McEwen's petition in error.

Within 30 days after the May order, McEwen filed a notice of appeal. In case No. A-17-638, the Court of Appeals summarily dismissed the appeal for lack of jurisdiction. The court's summary order explained that McEwen's motion for new trial did not "toll" the time to file a notice of appeal and that McEwen's notice of appeal was not timely filed.

McEwen moved for rehearing in the Court of Appeals. He argued that the May order was itself a final order. He premised this argument upon § 25-2001 and this court's decision in *Capitol Construction v. Skinner*.<sup>3</sup> Notably, McEwen \*\*557discussed two of our cases, *Goodman v. City of Omaha*<sup>4</sup> and *Timmerman v. Neth*,<sup>5</sup> regarding the applicability of a motion to alter or amend a judgment<sup>6</sup> where a district court acts as an intermediate appellate court. Based on this case law, McEwen conceded, as he did at oral argument before this court, that his motion to vacate did not act as a motion to alter or amend the judgment, which would have terminated the running of the appeal time.

By a summary order in case No. A-17-638, the Court of Appeals denied rehearing. The court explained that McEwen's motion to vacate did not "toll" the time to appeal from the March judgment. The court added, "Accordingly, by the time the district court entered [the May] order denying the motion to vacate, [McEwen] \*126could only appeal from that order. However, [McEwen's] brief on appeal argues only that the district court erred by denying his petition in error in the March [judgment]."<sup>7</sup> Thus, the Court of Appeals concluded that it lacked jurisdiction to consider the merits of the March judgment and left in place the dismissal of the appeal from the May order.

McEwen timely petitioned for further review, which we granted.<sup>7</sup>

After oral argument in this court, we requested supplemental briefing by the parties addressing whether, in light of the distinction traditionally recognized between petitions in error and appeals created by various statutes, the Legislature intended for motions to alter or amend a judgment under § 25-1329 to apply to judgments entered in error proceedings and, if so, the proper application of that statute to the case before us. The parties promptly submitted supplemental briefs, which we have considered.

#### \*\*558III. ASSIGNMENTS OF ERROR

McEwen assigns that the Court of Appeals erred in (1) dismissing his appeal for lack of jurisdiction and (2) overruling his subsequent motion for rehearing.

#### IV. STANDARD OF REVIEW

A jurisdictional question which does not involve a factual dispute is determined by an appellate court as a matter of law, which requires the appellate court to reach a conclusion independent of the lower court's decision.<sup>8</sup>

The meaning of a statute is a question of law.<sup>9</sup>

## C.2 GPT4o Deep Research Output (Excerpt)

Understood. I will research whether Neb. Rev. Stat. § 25-1329 applies to a district court's judgment resolving a petition in error, despite prior case law suggesting that it does not apply to district courts acting in an appellate capacity. I will craft the remaining portion of an appellate court judgment with legal analysis, discussion, and a reasoned conclusion, using only precedent and legal interpretations available up to July 12, 2019, and excluding McEwen v. Neb. State.

### ## ANALYSIS

#### ### District Court as Appellate Tribunal and Tolling Motions

Nebraska law has long held that when a district court exercises appellate jurisdiction – reviewing a decision of a lower court or tribunal – certain postjudgment motions are not available to extend the time for appeal. In particular, a motion for new trial is confined to trial courts; if the district court acted as an appellate court, a “motion for new trial” is not a proper pleading and does not arrest the running of the appeal clock. This rule has been applied uniformly whether the district court was hearing an appeal from a county court or from “some other lower tribunal”. Accordingly, our cases repeatedly refused to recognize such motions as tolling the time for appeal in that context. We reasoned that in an appellate review, the district court’s decision is not a true “judgment” as contemplated by the tolling statutes, but rather an appellate disposition of another body’s judgment. We thus defined “judgment” for purposes of a motion to alter or amend (§ 25-1329) as “the final determination of the rights of the parties in an action” – i.e. the decision of a trial court – and not an appellate decision of a district court. Under that view, a district court sitting as an intermediate appellate court could not entertain a tolling motion (whether styled as new trial or to alter/amend), and any such motion would not extend the 30-day deadline for filing a notice of appeal. This doctrine was well established in our jurisprudence.

#### ### Petition in Error Proceedings as Distinct from Appeals

Notwithstanding the above principles, we recognize a clear distinction in Nebraska law between a direct \*appeal\* to the district court and a \*proceeding in error\*. A proceeding in error in the district court (pursuant to Neb. Rev. Stat. §§ 25-1901 to 25-1908) is “distinct and independent”, whereas an appeal is “a mere continuation of the same cause in another court.”\* Unlike a direct appeal which removes the entire cause to the higher court for retrial on law and fact, an error proceeding is initiated by filing a petition in the district court (along with a transcript of the inferior tribunal’s record) and issuing summons to the adverse party. In other words, the petition in error functions as a new civil action filed in the district court, the purpose of which is to obtain review of the inferior body’s decision on the record. We have said that the two procedures – appeal and error – are so different that the statutory provisions governing each “cannot be taken together and construed as if they were one law”. Consistent with that view, the outcome of a petition in error is a judgment of the district court concluding the independent error proceeding, whereas on a direct appeal the district court’s ruling might simply remand the cause for further proceedings below. This historical distinction between an error proceeding and a regular appeal is crucial in assessing the applicability of § 25-1329.

Earlier Nebraska decisions did not always articulate this distinction when considering postjudgment motions. The broad rule stated in cases like *Goodman v. City of Omaha*\*\* (a zoning appeal) and *Timmerman v. Neth*\*\* (an administrative license appeal) was that no tolling motion lies when the district court acts as an appellate reviewer. Those cases, however, involved ordinary appeals where the district court was unquestionably sitting as an intermediate appellate court. We acknowledge that our jurisprudence at times used loose language referring to an “appeal by petition in error” and thus lumped error proceedings together with other appeals. Upon closer examination, however, the district court’s role in a petition in error is more akin to that of a trial court entering a judgment in an original action (albeit on a limited record review) than that of a purely intermediate appellate tribunal. The question, then, is whether the tolling provisions of § 25-1329 should apply to the judgment rendered by a district court at the conclusion of a petition in error proceeding. We conclude that they should.

#### ### Legislative Intent of § 25-1329

Section 25-1329 was enacted as part of L.B. 921 in 2000, against the backdrop of the above-described case law. By its plain terms, the statute imposes only two requirements for a motion to alter or amend: it must be filed within ten days after the entry of “the judgment,”\* and it must seek a substantive change to that judgment. Nothing in the text of § 25-1329 limits the term “judgment” to a particular court or type of proceeding. In contrast to other contexts where the Legislature has explicitly differentiated appeals from other actions, here the statute is written generally, suggesting it encompasses any final judgment properly so called. Indeed, the very purpose of § 25-1329, like similar provisions for new trial motions, is to give the court entering judgment a chance to correct any errors timely brought to its attention, thereby potentially obviating the need for an appeal. This purpose would seem equally valuable whether the “judgment” is reached after a full trial or upon a review of another tribunal’s decision. We must presume the Legislature intended § 25-1329 to be construed in a manner that effectuates this remedial purpose, rather than in a way that creates procedural traps or unnecessary appeals.

The legislative history of L.B. 921 confirms that § 25-1329 was designed to fill a procedural gap. Before 2000, a party aggrieved by a decision that was not the result of a trial (for example, a summary judgment or dismissal before trial) faced uncertainty about how to seek reconsideration. A Judiciary Committee comment on the 2000 amendments explained that a motion to alter or amend was intended as a mechanism to seek relief from “a decision based on something other than a trial.”\* In the same bill, the new trial statutes were clarified to refer to a verdict or trial decision, underscoring that a “new trial” motion is inappropriate when no trial occurred. Thus, the Legislature provided that in such instances, a motion to alter or amend the judgment under § 25-1329 would be the proper means to ask the court to reconsider its decision. We find no indication that the lawmakers intended to withhold this means of relief in the subset of cases where the district court’s judgment happened to arise from an error proceeding. On the contrary, error proceedings fit neatly into the very category of decisions “based on something other than a trial” that § 25-1329 was meant to address. Had the Legislature wanted to exclude appellate-type district court judgments from § 25-1329, it could have said so explicitly. Instead, the statute’s broad language and purpose signal an intent to treat any final judgment of the district court – including one rendered at the conclusion of a petition in error – as eligible for a timely motion to alter or amend. This interpretation best harmonizes the statute with its goal of avoiding needless delay and expense from premature appeals.

#### ### Application to the Present Case

In light of the foregoing, we hold that Neb. Rev. Stat. § 25-1329 applies to a district court’s judgment resolving a petition in error. When a party timely files a motion to alter or amend such a judgment (seeking a substantive change in the outcome), the motion terminates the running of the time for appeal,....

## D Kendall correlation for human-LLM agent evaluation

Table S1: Meta-evaluation on citation relevance (CT), constraint extraction (CS), and argument validity (AG) across various perturbation settings. **Kendall correlations** are reported, in percentage. Values are colored by the blue-to-red coolwarm colormap.

| o3                          |    |     | Claude |     |     | G-Pro |    |     | G-Flash |     |     | Qwen2.5-72B |     |     |    |
|-----------------------------|----|-----|--------|-----|-----|-------|----|-----|---------|-----|-----|-------------|-----|-----|----|
|                             | CT | CS  | AG     | CT  | CS  | AG    | CT | CS  | AG      | CT  | CS  | AG          | CT  | CS  | AG |
| <b>GPT-4o Deep Research</b> |    |     |        |     |     |       |    |     |         |     |     |             |     |     |    |
| default                     | 28 | 19  | 18     | 29  | 15  | 0     | 10 | 20  | 5       | 14  | 0   | 3           | 5   | 22  | 27 |
| most important              | 39 | 50  | 0      | 43  | 47  | -2    | 4  | -13 | -5      | 3   | 5   | -10         | 40  | 18  | 21 |
| 2nd most important          | 29 | 9   | 38     | 13  | 2   | 6     | 41 | 0   | 19      | 3   | 31  | 12          | 0   | 39  | 26 |
| least important             | 18 | -28 | 0      | 14  | 6   | -10   | 15 | 12  | 28      | -12 | -8  | 17          | 12  | -8  | 4  |
| opposite                    | 33 | 14  | 13     | 49  | 64  | 67    | 3  | 36  | 25      | 35  | 37  | 25          | 41  | 50  | 50 |
| <b>Gemini Deep Research</b> |    |     |        |     |     |       |    |     |         |     |     |             |     |     |    |
| most important              | 47 | 48  | 44     | 20  | 32  | 29    | 45 | 46  | 5       | 39  | 44  | 32          | 9   | 23  | 18 |
| 2nd most important          | 16 | 9   | 11     | 25  | 24  | 25    | 40 | 31  | 11      | 6   | 28  | 31          | 6   | -20 | 8  |
| 2nd most important          | 93 | 0   | -31    | 25  | 24  | 25    | 40 | 31  | 11      | 6   | 28  | 31          | 6   | -20 | 8  |
| <b>o3-search</b>            |    |     |        |     |     |       |    |     |         |     |     |             |     |     |    |
| most important              | 31 | 23  | -7     | 35  | 27  | 14    | -2 | -1  | -9      | 0   | 0   | 0           | 0   | 20  | 0  |
| 2nd most important          | 6  | 45  | 6      | 5   | -7  | -4    | 3  | 6   | 12      | -3  | -11 | -28         | 18  | -33 | -1 |
| <b>open-deep-research</b>   |    |     |        |     |     |       |    |     |         |     |     |             |     |     |    |
| most important              | -3 | 0   | 9      | 32  | 23  | 39    | 20 | 38  | 66      | 7   | 16  | 49          | 36  | 23  | 35 |
| 2nd most important          | 0  | 19  | 33     | 42  | 9   | 6     | 6  | 16  | 3       | 6   | 16  | 49          | -10 | 0   | 20 |
| <b>OpenRag</b>              |    |     |        |     |     |       |    |     |         |     |     |             |     |     |    |
| default                     | 0  | 80  | 0      | -55 | 38  | 4     | 0  | 59  | 27      | 0   | 4   | 25          | 0   | 0   | 0  |
| most important              | 0  | 0   | 26     | -25 | -15 | -26   | 0  | 49  | 12      | -17 | 0   | -25         | 0   | 0   | 0  |
| 2nd most important          | 0  | 37  | 0      | 40  | 4   | 0     | 0  | 41  | 7       | 0   | 22  | -7          | -41 | -58 | 0  |