CourtReasoner: Can LLM Agents Reason Like Judges?

Anonymous authors
Paper under double-blind review

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 for evaluating LLM agents' capabilities in complex legal reasoning. Sourcing U.S. court opinions, we construct benchmarks that measure the LLMs ability to construct goal-oriented legal reasoning. CourtReasoner measured the agent's ability to argue both ways in a legal dispute, rather than simple Q/A. Our results show that more than 60% of frontier model outputs contain invalid arguments and more than 53% of frontier model outputs 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. We will release our data, code and full annotation guidelines publicly for future research.

Large Language Models (LLMs) have achieved strong performance across mathematical, logical, and commonsense reasoning tasks (OpenAI, 2025; OpenAI; DeepMind, 2025; Anthropic, 2024). However, legal reasoning in the U.S. legal system presents distinct challenges. Unlike math problems with a single correct answer, legal cases often involve ambiguity, multiple plausible outcomes, and frequent expert disagreement—especially in appellate or litigated cases (Post, 2024). Metrics like binary accuracy, used in prior legal QA benchmarks (Guha et al., 2023; Koreeda & Manning, 2021; Hendrycks et al., 2021), fail to capture the nuanced reasoning required in law.

Legal reasoning demands more than factual recall. It involves applying doctrine to complex facts, reasoning through precedent, identifying competing interpretations, and constructing persuasive legal arguments (Schauer, 2009; Hanks et al., 1994; Dworkin, 1986). Evaluation must therefore consider how well a model engages with legal texts, differentiates counterarguments, and constructs logically coherent and persuasive narratives. This is compounded by ultra-long context windows, with inputs and outputs often exceeding tens of thousands of tokens (Zheng et al., 2025). Simple test-time scaling of input/output lengths (Muennighoff et al., 2025) is insufficient and computationally infeasible for such tasks.

To tackle this challenge, we introduce CourtReasoner, a benchmark derived from real U.S. court opinions from `https://case.law/caselaw/`. Unlike bar exam-style or multiple-choice tasks, CourtReasoner evaluates full-length judicial reasoning. Each document is segmented into "background and facts" and "reasoning" sections. The model must generate the reasoning portion, which involves breaking the issue into doctrinal components, citing one precedent per component, synthesizing sub-conclusions, addressing outlier precedents, and producing a logically structured conclusion. This process may repeat for multiple doctrines—e.g., standing, immunity, merits—before reaching a final decision.

To prevent models from copying known text (given all cases are publicly accessible online), we create three adversarial variants. Human annotators redact or alter facts to present novel scenarios, testing whether LLMs can formulate new arguments rather than rely on memorized or retrieved content. These adversarial settings preserve the logical structure of the case while withholding key factual triggers seen during pretraining.

We collaborate with experienced U.S. legal professionals to create meta-evaluation for model-generated outputs along three expert-validated axes: (1) Citation Relevance — is the precedent correctly chosen and applied? (2) Constraint Extraction — are the necessary conditions for applying the cited rule correctly identified? (3) Argument Validity — do conclusions follow logically from the rules and facts?

We evaluate several cutting-edge LLMs, including GPT-4o Deep Research (OpenAI), OpenAI `o3`(OpenAI, 2025), Gemini Pro Deep Research(Google, 2024), and open-source baselines such as Open Deep Research[1]. Our meta-evaluation benchmark has an inter-annotator agreement of 75% or higher across all model settings. While models show emerging capacity for structured legal reasoning, they often fall short in citation accuracy, constraint comprehension, and argument soundness—especially when required to argue for a legally plausible but opposite position. For instance, citation relevance drops from 60% when aligned with the original court's stance to 30% in adversarial cases.

We also explore LLMs as evaluators. OpenAI `o3` and Claude-3.7 (Anthropic) achieve moderate correlation with human evaluation in some settings, but no model consistently performs well across all meta-evaluation tasks, including Gemini (Google DeepMind, 2025), Gemini-Flash (DeepMind, 2025), and Qwen-2.5-72B (Yang et al., 2024).

In summary, our contributions are: (1) a realistic benchmark, CourtReasoner, capturing complex, full-length legal reasoning; (2) expert-defined evaluation axes — citation relevance, constraint extraction, and argument validity; and (3) quantitative and qualitative analyses of model capabilities and failures, including their limits in moral reasoning under legal contexts.

1 Dataset Construction

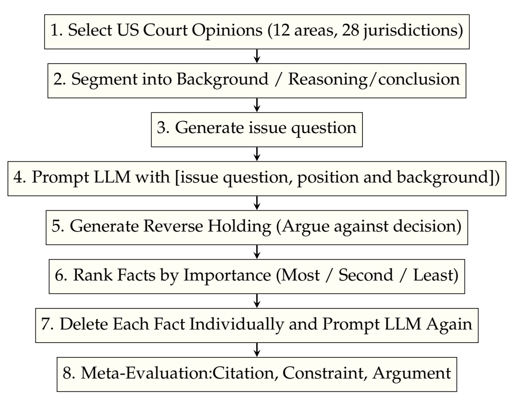

Figure 1: Data construction for COURTREASONER.

The annotation process begins with selecting diverse U.S. appellate court opinions from `https://case.law`, prioritizing variability in jurisdiction and reasoning complexity. Annotators segment each opinion by identifying a sub-section's opening paragraph, which typically introduces the legal question. Using GPT-4o, this paragraph is converted into a clear, well-formed legal issue, with the remaining section serving as the model's target output. Issues are manually reviewed and edited for legal correctness.

Each model is given a prompt comprising the generated legal question, the Introduction, Background, Facts, and the side it should argue for. Models used include Gemini Pro Deep Research, GPT-4o with Deep Research, OpenAI `o3` with search, and Open Deep Research [2].

[1] https://github.com/langchain-ai/open_deep_research
[2] `https://github.com/langchain-ai/open_deep_research`

Table 1: 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-deep-research		Open-Rag		
	most	2nd most	east	default	most	2nd most	least	opposite	most	2nd most	most	2nd most	default	most	2nd most
citation relevance = 0	13	10	20	0	0	0	0	17	0	0	0	0	0	0	0
citation relevance = 1	13	7	0	0	6	0	6	8	4	4	30	0	0	0	0
citation relevance = 2	30	41	20	4	24	25	12	17	12	19	30	40	10	20	20
citation relevance = 3	40	28	0	58	35	44	44	33	73	62	40	50	80	80	70
citation relevance = 4	3	14	60	38	35	31	38	25	12	15	0	10	10	0	10
constraint extraction = 0	7	3	40	0	0	0	0	17	0	0	0	0	0	0	0
constraint extraction = 1	10	14	0	4	6	0	6	17	0	4	20	0	0	0	0
constraint extraction = 2	47	28	0	8	47	56	44	0	42	46	30	20	30	60	30
constraint extraction = 3	20	34	0	75	24	19	19	67	31	38	20	40	40	20	40
constraint extraction = 4	17	21	60	12	24	25	31	0	27	12	30	40	40	20	30
argument validity = 0	3	3	20	0	0	0	0	17	0	0	0	0	0	0	0
argument validity = 1	27	34	20	0	0	12	19	8	4	4	20	0	0	0	0
argument validity = 2	37	21	0	12	53	44	25	8	42	46	30	40	30	60	60
argument validity = 3	17	24	0	67	29	12	25	67	46	50	40	50	40	30	30
argument validity = 4	17	17	60	21	18	31	31	0	8	0	10	10	30	10	10

Additionally, the authors introduce Legal OpenRAG, a retrieval-augmented generation framework that includes: (1) an E5-base embedding-based retriever indexing 6.7M case law documents (Wang et al., 2022), (2) a GPT-4o reranker, and (3) an o3-based generator that uses the top 10 reranked precedents.

Model outputs are then evaluated through structured error analysis. Annotators assess each cited case for (A) relevance (high/remote/irrelevant), (B) completeness of extracted legal constraints, and (C) whether reasoning convincingly shows constraint satisfaction. Errors in logic, irrelevant citations, and "hand-wave" arguments are flagged. High-quality citations that streamline argumentation are tagged as “good.” The full rubric is detailed in the appendix.

To probe genuine legal reasoning, the adversarial evaluation removes key facts from known cases. Annotators manually rank each fact by importance to the original court’s conclusion. Three versions of the case are created by removing one highly relevant, one moderately relevant, and one trivial fact. GPT-4o is then prompted to construct new, logically sound arguments without reusing the original case’s methodology, citing only older precedents and avoiding the deleted premise. This setup stress-tests the model’s ability to reason, not memorize.

1.1 Dataset statistics

Task settings and models We compiled 319 expert-annotated meta-evaluation examples from 50 seed cases to compare five LLM variants under diverse conditions. Gemini-Deep-Research was assessed with two fact-deletion adversaries (31 examples each, 62 total). GPT-4o-Deep-Research covered five settings: original same-side (50), opposite-side (50), plus three citation-deletion adversaries (15 each), totaling 145. o3-search underwent the two deletion variants, adding 52 examples. Open-deep-research and Open-RAG each contributed 30 examples—original plus two adversarial variants (10 per setting).

This design probes robustness to factual perturbations and differing retrieval-generation workflows. Default input questions average 1,758 words and 71 sentences; deleting a fact trims words slightly while sentence count stays steady.

Output lengths diverge sharply. Gemini-Deep-Research is the most verbose, averaging 5,023 words and 184 sentences. GPT-4o-Deep-Research is shorter yet detailed at 2,614 words and 91 sentences. Open-RAG and open-deep-research sit mid-range, averaging 2,094 and 1,635 words. o3-search is the briefest, with 1,060 words and 46 sentences, suggesting a minimal or depth-limited style. These contrasts highlight substantial variability in reasoning verbosity and strategy across LLMs.

Meta-evaluation Table 1 presents human evaluation scores for model-generated legal reasoning across citation relevance, constraint extraction, and argument validity. Scores range from 0 to 4, with darker shades indicating stronger performance. The results show a consistent trend: performance degrades when key facts—especially the most important ones—are removed.

Gemini Deep Research is most affected, showing increased low scores and reduced high ones. For example, with the most important fact deleted, 13% of its outputs score 0 in citation relevance, and only 3% score 4.

GPT-4o Deep Research shows more robustness, maintaining high ratings (3–4) even with key fact deletions. This likely reflects stronger reasoning and better retrieval reranking. However, its performance still drops in the "opposite-side" setting, especially in citation relevance and constraint extraction.

Open-RAG performs well by default, with 80% of outputs rated 3 or 4 in citation relevance and no low scores in the other dimensions. Its design—limiting retrieval to case law and separating retrieval from generation—seems to enhance reasoning quality. Still, performance declines when important facts are removed.

By contrast, o3-search and open-deep-research produce flatter score distributions, suggesting weaker citation grounding. These results highlight that while better models show promise, current LLMs remain brittle under adversarial perturbations.

2 Experiments and Analysis

2.1 LLM as an Evaluator

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, 2025), Claude-3.7 (Anthropic), Gemini-Pro (Google DeepMind, 2025), Gemini-Flash (DeepMind, 2025) and Qwen2.5-72B (Yang et al., 2024).

Each evaluator receives (i) the question containing the position the model is asked to take and the background and introduction sections of the case, (ii) the candidate analysis, and (iii) the three-layer rubric for *citation relevance*, *constraint extraction*, and *argument validity* that human annotators already follow.

Correlation with human evaluation Spearman and Kendall correlations between LLM-based graders and human annotations across citation relevance (CT), constraint extraction (CS), and argument validity (AG) appear in the Appendix. Claude aligns best with human judgments, especially in adversarial settings. On GPT-4o's "opposite-side" outputs it reaches 75/64 (CS) and 81/67 (AG), showing strong sensitivity to flipped reasoning and constraint breaches.

Qwen2.5-72B is similarly robust on GPT-4o and Gemini outputs but correlations often fall to zero or negative on OpenRAG and open-deep-research, exposing challenges in grading retrieval-augmented text. Gemini Pro and Gemini Flash evaluate their own generations well yet drop sharply on other models, indicating limited generalization. OpenAI's o3 correlates moderately with its own and Gemini outputs but deteriorates under fact-deletion, sometimes turning negative. Overall, grader reliability varies widely across models and perturbations. Claude and Qwen2.5 are promising, but stable meta-evaluation may require ensemble or model-aware scoring. Means ± standard deviations are provided in the Appendix.

3 Analysis

3.1 Granular error categories

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

Table 2: 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.

In evaluating the legal-reasoning capabilities of large language models (LLMs), we uncovered recurring errors that compromise analytical reliability. Expert annotators produced free-form analyses, which we parsed into granular categories (Table 2). A common failure mode is that LLMs handwave when they cannot derive a valid solution, asserting conclusions without a coherent chain. This limitation shows that frontier models still struggle with the complexity of structured legal reasoning. One clear manifestation is factual errors: models misstate case holdings, procedural histories, or material facts, undermining their analysis.

Another frequent problem is quoting without citation, which obscures traceability. Models may also cite sources but quote inaccurately, or quote correctly yet cite incorrectly, distorting legal authority. Some responses give incomplete analyses that omit key factors, while others perform unnecessary analysis of tangential issues. A final category, change in legal stance, appears when the model shifts to an unintended party's viewpoint. Representative case studies follow.

Exaggeration or change of key phrases In State v. Kony[3], GPT-4o Deep Research inserted the phrase "vast majority" into Dr. Bivens's testimony, which merely stated that every incest offender acted in the home. The court labeled the evidence "significantly misleading," not "very high" in probative value.

Analyzing an issue not legally reviewable In Visser v. Auto Alley, LLC[4], o3-search ignored the rule that stipulated judgments are not appealable. It spent time evaluating enforceability, overlooking that such judgments may be challenged only under exceptional conditions like fraud.

Quoting without citation In Fischer v. City of Sioux Falls[5], GPT-4o Deep Research stated that "gross negligence and willful or wanton misconduct mean the same thing in South Dakota" without citing precedent. Although substantively correct, the claim appears as an unsupported assertion rather than a principle grounded in authority.

3.2 Qualitative analysis on interleaving legal and normative reasoning

Unlike purely deductive domains such as math or logic, legal reasoning often involves normative judgment. Vague statutory terms, ambiguous language, and undeveloped precedent compel judges to invoke moral reasoning. One example in COURTREASONER concerns whether a statute banning "substantial" emotional distress is First-Amendment compliant; another asks if a dwelling's disrepair was assessed "reasonably." These tasks highlight the tension between descriptive precedent and prescriptive morality that pervades common-law adjudication. When statutory language is indeterminate, courts turn to policy considerations, social norms, and constitutional principles to anchor their reasoning. LLMs must therefore learn not only to retrieve controlling authority but also to weigh extra-textual values—a capability current models lack. Bridging this gap will require training signals that reward accurate citation while penalizing moral overreach, plus datasets pairing doctrinal analysis with explicit normative justification. These challenges remain largely unsolved for practitioners and system designers alike.

4 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. 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.

[3] https://case.law/caselaw/?reporter=hawvolume=138case=0001-01

[4] https://case.law/caselaw/?reporter=idahovolume=162case=0001-01

[5] https://case.law/caselaw/?reporter=nw2dvolume=919case=0211-01

References

Anthropic. Claude 3.7 sonnet and claude code. URL https://www.anthropic.com/news/claude-3-7-sonnet. Announcement blog post.

Anthropic. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet, June 2024.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. Can gpt-3 perform statutory reasoning?, 2023. URL https://arxiv.org/abs/2302.06100.

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

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

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

Ronald Dworkin. *Law's Empire*. Harvard University Press, 1986. ISBN 978-0-674-51836-0. Google-Books-ID: 8TCx_UbbI3AC.

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

Google DeepMind. Gemini pro, 2025. URL https://deepmind.google/technologies/gemini/pro/. Accessed: 2025-05-20.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models, 2023.

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

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*, 2021.

Nils Holzenberger and Benjamin Van Durme. Factoring statutory reasoning as language understanding challenges. *arXiv preprint arXiv:2105.07903*, 2021.

Yuta Koreeda and Christopher D Manning. Contractnli: A dataset for document-level natural language inference for contracts. *arXiv preprint arXiv:2110.01799*, 2021.

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

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling, 2025. URL https://arxiv.org/abs/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. Large language models as tax attorneys: A case study in legal capabilities emergence, 2023. URL https://arxiv.org/abs/2306.07075.

OpenAI. Introducing deep research. URL https://openai.com/index/introducing-deep-research/. Announcement blog post.

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

Robert Post. The Supreme Court's Crisis of Authority: Law, Politics, and the Judiciary Act of 1925, December 2024. URL https://papers.ssrn.com/abstract=5075524.

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

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

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

Daniel Schwarcz, Sam Manning, Patrick James Barry, David R. Cleveland, J. J. Prescott, and Beverly Rich. Ai-powered lawyering: Ai reasoning models, retrieval augmented generation, and the future of legal practice. Technical Report 25-16, Minnesota Legal Studies Research Paper, March 2025. 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. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

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

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. 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)*, pp. 1330–1340, 2016.

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. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. When does pretraining 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*, pp. 159–168, 2021.

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D. Manning, Peter Henderson, and Daniel E. Ho. A Reasoning-Focused Legal Retrieval Benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, CSLAW '25, pp. 169–193, New York, NY, USA, March 2025. Association for Computing Machinery. ISBN 9798400714214. doi: 10.1145/3709025.3712219. URL https://dl.acm.org/doi/10.1145/3709025.3712219.

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

Table 3: 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

A Related work

A.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. (2023) 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 & Manning, 2021; Hendrycks et al., 2021; Wang et al., 2023; Wilson et al., 2016; Zheng et al., 2021; Zimmeck et al., 2019; Ravichander et al., 2019; Holzenberger & 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.

A.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). 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 & 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).

B Input/output length

Input/output length We summarize the distribution of input questions and output model generations in Table 3 and Table 4 respectively.

Table 4: 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

Table 5: Performance of LLMs evaluated by human annotators, reported in mean$_{\text{(standard deviation)}}$.

	Gemini Deep Research			GPT-4o Deep Research					o3-search		open-deep-research		Open-Rag		
	× most	× 2nd most	× least	default	× most	× 2nd most	× least	opposite	× most	× 2nd most	× most	× 2nd most	default	× most	× 2nd most
citation relevance	2.07$_{(1.09)}$	2.28$_{(1.11)}$	2.80$_{(1.60)}$	3.33$_{(0.55)}$	3.00$_{(0.91)}$	3.06$_{(0.75)}$	3.13$_{(0.86)}$	2.42$_{(1.38)}$	2.92$_{(0.62)}$	2.89$_{(0.70)}$	2.10$_{(0.83)}$	2.70$_{(0.64)}$	3.00$_{(0.45)}$	2.80$_{(0.40)}$	2.90$_{(0.55)}$
constraint extraction	2.30$_{(1.07)}$	2.55$_{(1.07)}$	2.40$_{(1.96)}$	2.96$_{(0.61)}$	2.65$_{(0.90)}$	2.69$_{(0.85)}$	2.75$_{(0.97)}$	2.17$_{(1.21)}$	2.85$_{(0.82)}$	2.58$_{(0.74)}$	2.60$_{(1.11)}$	3.20$_{(0.75)}$	3.10$_{(0.83)}$	2.60$_{(0.80)}$	3.00$_{(0.78)}$
argument validity	2.17$_{(1.10)}$	2.17$_{(1.18)}$	2.60$_{(1.74)}$	3.08$_{(0.57)}$	2.65$_{(0.76)}$	2.63$_{(1.05)}$	2.69$_{(1.10)}$	2.25$_{(1.16)}$	2.58$_{(0.69)}$	2.46$_{(0.57)}$	2.40$_{(0.92)}$	2.70$_{(0.64)}$	3.00$_{(0.78)}$	2.50$_{(0.67)}$	2.50$_{(0.67)}$

C Performance of LLMs reported in mean

Table 5 reports human evaluation scores of each model's legal reasoning performance across three key dimensions: citation relevance, constraint extraction accuracy, and argument validity. Scores are presented as means with standard deviations. Each model was tested under a range of perturbation settings, including deletion of the most, second most, and least important facts, as well as an "opposite side" adversarial 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 perturbed settings suggests that GPT-4o benefits from both strong citation grounding and stable logical reasoning. We hypothesize that its consistently high performance stems from improved fine-tuning on legal domain research tasks and a reranking mechanism that helps filter out less relevant precedents during retrieval. 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.

D 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>",
}}
```
~~~

E Example of model input and output

E.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 .3 Notably, McEwen **557discussed two of our cases, Goodman v. City of Omaha4 and Timmerman v. Neth ,5 regarding the applicability of a motion to alter or amend a judgment6 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]." 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.7

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.8

The meaning of a statute is a question of law.9

E.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,...

E.3 Spearman and Kendall correlation for human-LLM agent evaluation

Table 6: Meta-evaluation on citation relevance (CT), constraint extraction (CS), and argument validity (AG) across various perturbation settings. **Spearman correlations** are reported (percentage).

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

Table 7: Meta-evaluation on citation relevance (CT), constraint extraction (CS), and argument validity (AG) across various perturbation settings. **Kendall correlations** are reported (percentage).

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