Investigating the Shortcomings of LLMs in Step-by-Step Legal Reasoning

Anonymous ACL submission

Abstract

Reasoning abilities of LLMs have been a key focus in recent years. One challenging reasoning domain with interesting nuances is legal reasoning, which requires careful application of rules, and precedents while balancing deductive and analogical reasoning, and conflicts between rules. Although there have been a few works on using LLMs for legal reasoning, their focus has been on overall accuracy. In this paper, we dig deeper to do a step-by-step analysis and figure out where they commit errors. We use the college-level Multiple Choice Question-Answering (MCQA) task from the *Civil Procedure* dataset and propose a new error taxonomy derived from initial manual analysis of reasoning chains with respect to several LLMs, including two objective measures: soundness and correctness scores. We then develop an LLM-based automated evaluation framework to identify reasoning errors and evaluate the performance of LLMs. The computation of soundness and correctness on the dataset using the auto-evaluator framework reveals several interesting insights. Furthermore, we show that incorporating the error taxonomy as feedback in popular prompting techniques marginally increases LLM performance. Our work will also serve as an evaluation framework that can be used in detailed error analysis of reasoning chains for logic-intensive complex tasks[1].

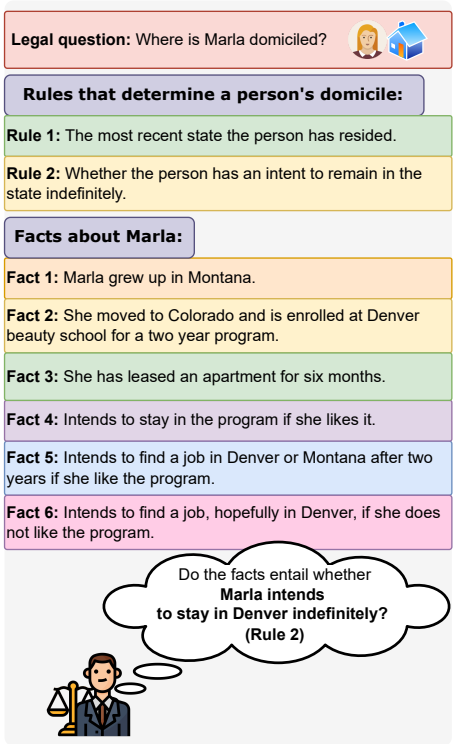

Figure 1: An example of determining domicile in a legal context. A reasoner must discern whether the condition of 'indefinite to stay in a place' is met. While many LLMs predict Marla is domiciled in Montana since her program is only for 2 years, legally, her ambiguous plans indicate an intent to remain in Denver indefinitely, making her domiciled in Denver, not Montana.

1 Introduction

Legal reasoning is a complex process requiring the careful application of rules, and precedents while balancing deductive and analogical reasoning, for various legal scenarios (Walker, 2007; Hafner and Berman, 2002). These challenges are heightened by reasoning through uncertainties (Figure 1) and ambiguous laws (Figure 18 (Appendix J)). In recent years, Large Language Models (LLMs) have emerged as the most dominant AI models to process and generate natural language. There has been widespread research to showcase the emergence of natural language understanding (NLU) and reasoning abilities (Zelikman et al., 2022, 2024; Hao et al., 2023; Mondorf and Plank, 2024; Lanchantin et al., 2024) of LLMs. These findings have opened up avenues to utilize LLMs in complex domains like Law (Lai et al., 2023). Consequently, there has been significant research in evaluating the performance of LLMs' legal reasoning ability (Guha

[1]Data and source code are available at https://anonymous.4open.science/r/legal_reasoning-D5F6/

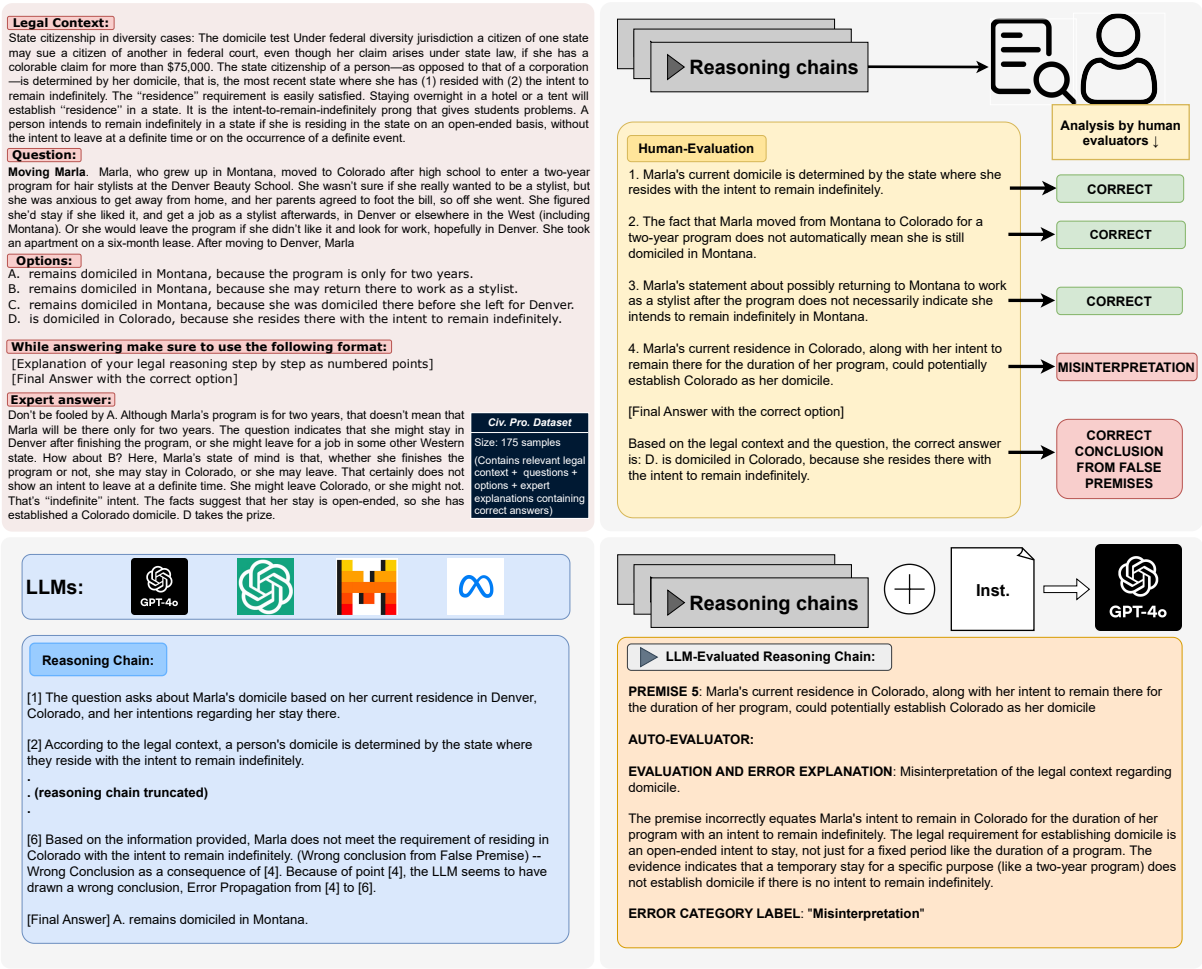

Figure 2: Overview of the proposed pipeline for evaluating legal reasoning in LLMs. The process begins with converting *Civ. Pro.* dataset (**top left**), followed by generating reasoning chains using LLMs in a zero-shot CoT setting (**bottom left**). These chains are manually analyzed for various error types (**top right**), based on the proposed error taxonomy. The pipeline is then automated using an LLM-based system (**bottom right**) to assess reasoning chains for errors such as misinterpretation, providing insights into the LLMs' reasoning accuracy.

et al., 2023; Blair-Stanek et al., 2023; Kang et al., 2023). *In this paper, our main aim is to do a fine-grained analysis of errors in step-by-step legal reasoning.* Earlier works of step-by-step reasoning in (Golovneva et al., 2023; Prasad et al., 2023) do not specifically cater to legal reasoning.

As shown in Figure 1, analyzing legal scenarios requires extensive consideration of critical analysis of prior context. Hence, beyond just evaluating final answers, it is crucial to analyze the step-by-step reasoning chains generated by LLMs and where they falter to gauge their reasoning capabilities to solve such tasks. Despite continued improvements, LLM outputs are affected by fundamental challenges such as hallucinations (Dahl et al., 2024), and misunderstanding long-contexts (Lu et al., 2024). Our work aims to investigate such errors on a fine-grained level for each step generated to showcase the ability of LLM to perform legal reasoning. To this end, we leveraged the dataset from Bongard et al. (2022) (referred to as '*Civ. Pro.*' throughout the paper) in the form of MCQA and analyzed reasoning chains generated by LLMs. The dataset comprises of 175 legal scenarios, each providing a comprehensive legal context that includes relevant rules, precedents, and exceptions. Following this context, questions and options are presented to test the understanding and application of these legal principles. The primary goal of using this dataset is to assess how well the LLM reasons logically and contextually when provided with all necessary legal rules, without relying on its inherent knowledge to recall them. This approach highlights the LLM's ability to perform deductive reasoning in complex, context-rich legal scenarios. The most recent and widely used LLMs,

Mistral-7B, Llama-3-8B, GPT-3.5-turbo, GPT-4-turbo, and GPT-4o, are evaluated in zero-shot-CoT setting (Wei et al., 2023) (as shown in Figure 2).

Subsequently, human evaluation of the LLM-generated reasoning chains against the ground truth solutions was performed for a subset of the dataset to find the most common errors being committed in the reasoning steps. This resulted in the development of a detailed error taxonomy, as outlined in Tables 1 and 2, which allowed for a deeper understanding of the underlying causes of LLMs' failures. Additionally, to extend this evaluation across the entire dataset, we develop a method employing 'LLMs as auto-evaluators' (inspired by Liu et al. (2023); Chern et al. (2024)) that leverages our proposed error taxonomy. The auto-evaluator framework achieves a recall of $\sim$85.3% in terms of identifying errors similar to human annotation. To derive better qualitative observations, we utilize two LLM-based metrics: soundness and correctness (details in §4.2) to further analyze the reasoning chains. These metrics provide us with interesting insights, the most prominent of them being the inability to generate error-free rationales due to misinterpretation of the contextual nuances.

In the end, we apply various prompting strategies (details in §5.3) known to enhance LLM reasoning abilities with the zero-shot method. We also incorporated our error taxonomy as feedback to the prompting strategies to assess their effectiveness in mitigating errors in legal reasoning. The prompting strategies enhanced with error taxonomy feedback show improved accuracy (maximum of $\sim$4%) which suggests that providing information about errors help in improving the performance of LLMs. We hope that this framework for automatically evaluating step-by-step reasoning in complex tasks will be helpful for future research. In summary, the main contributions of our paper are:

1. We propose an error taxonomy to systematically identify the most commonly occurring errors in step-by-step legal reasoning.
2. We develop an LLM-based pipeline to automatically detect errors, and introduce two key metrics—soundness and correctness—to evaluate step-by-step legal reasoning.
3. We investigate the integration of error-taxonomy feedback into various LLM prompting strategies and find that it enhances the LLMs' legal reasoning capabilities.

2 Related Work

Legal Reasoning with LLMs LegalBench (Guha et al., 2023), LawBench (Fei et al., 2023), LEXGLUE (Chalkidis et al., 2021) and LEGALSEMI (Kang et al., 2024) have introduced exhaustive benchmarks which cover distinct tasks to measure the legal reasoning abilities of LLMs. Dahl et al. (2024) addresses the problem of hallucinations in LLMs, particularly their behavior in generating information that lacks factual accuracy. While these works have evaluated legal reasoning of LLMs on labelled predictions or through expert manual evaluation, our work specifically focuses on evaluating natural language step-by-step rationales with the help of LLMs. We also introduce a task-specific fine-grained error taxonomy to assess the reliability of LLMs in producing error-free reasoning chains.

Evaluation and Verification of Step-by-Step Reasoning Chains through LLMs ROSCOE (Golovneva et al., 2023) offers a comprehensive suite of metrics to assess various aspects of reasoning quality, including correctness, informativeness, consistency, and coherence. ReCEval (Prasad et al., 2023), on the other hand, specifically targets incorrect answer detection by specifically analyzing the 'correctness' and 'informativeness' of reasoning steps. (Ling et al., 2023). LLM Reasoners (Hao et al., 2024) introduces a novel evaluation framework for a detailed analysis of large language models' step-by-step reasoning abilities. Ling et al. (2023) introduces Natural Program, a natural language-based deductive reasoning format that decomposes a reasoning verification process into a series of step-by-step subprocesses. Tyen et al. (2024) explores the limitations of LLMs in detecting reasoning errors and highlights their effectiveness in correcting errors when provided with specific locations. Li et al. (2023) presents the DIVERSE approach involving a three-stage process to identify and correct errors at each step of the reasoning chain. While prior works introduce valuable evaluation frameworks, they are not optimized specifically for a legal reasoning task. Our work utilizes soundness and correctness metrics, which are simple yet effective in offering detailed insights into step-by-step legal reasoning errors. These metrics also make evaluation scalable and reduce the need for manual effort with our formulated LLM-based evaluation framework.

Category	Description
Misinterpretation	The LLM misinterprets or omits some part/entirety of the legal context, question or the options (or a combination of them). This usually leads to the wrong reasoning and selection of wrong conclusion. The following error instances fall under the taxon of misinterpretation: 1. Misunderstanding the legal rules. 2. Misunderstanding the legal situation/issue at hand. 3. Omission of parts of the provided context while reasoning. 4. Incompletely applying a legal rule. 5. Incorrectly applying the legal rule. 6. Wrong assumptions derived from the the provided context.
Irrelevant Premise	The LLM generates a premise which is not relevant in solving the question or that it may divert the reasoning chain from solving the question correctly. An premise may be logically valid and factually true but the absence of this premise can still lead to the correct conclusion.
Factual Hallucination	This error category covers instances where the LLM, during its reasoning process, generates information that is either inconsistent with the facts of the given legal scenario or is entirely fabricated with no basis in reality.

Table 1: Error taxonomy for the Premise-level steps. The taxonomy has been developed with consideration for the types of errors that a human reasoner might commit when constructing a rationale for a given legal scenario.

3 Evaluation of Reasoning Chains

3.1 The *Civil Procedure* Dataset

The dataset has been sourced from MCQs present in the "The Glannon Guide To Civil Procedure" (Glannon, 2019). We compiled the *Civ. Pro.* dataset with 175 samples of college-level law questions from the US Civil Procedure laws. The dataset includes relevant legal context, multiple-choice questions, and expert answers with correct explanations provided by legal experts. These elements were extracted and converted into a prompt-based format suitable for LLM inference and the generation of reasoning chains. The *Civ. Pro.* dataset consists of samples comprising of $\mathcal{D} = < lc_n, q_n, o_n, e_n >$, where lc_n, q_n, o_n and e_n denote the nth legal context, question, option-set and expert-answer respectively.

3.2 Manual Evaluation

Human evaluators are instructed to find flaws in a reasoning chain and explain the flaws in natural descriptive language. To create an error taxonomy, we adopt an exhaustive approach, continuously updating the taxonomy until no new errors are identified. Specifically, 120 reasoning chains were used for evaluation (generated as responses by four LLMs to the same 30 data sample subset). This evaluation helped to solidify our proposed taxonomy as described in §3.3. Further details regarding annotation guidelines and examples are provided in Appendix H and Appendix I (Tables 7-11).

3.3 Proposed Error Taxonomy

The error taxonomy is designed to mirror the types of errors humans make when reasoning about passage comprehension and constructing rational arguments. It classifies errors into two levels: 1. Premise-level and 2. Conclusion-level errors. While premise-level errors often influence errors at the conclusion level, many conclusion-level errors occur independently. Conclusion-level errors serve as indicators of the LLMs' overall accuracy in selecting the final answer to a legal question.

Premise-level Errors These errors have occurred in one of the premises of the reasoning chain. They highlight the core issue with LLMs ineffectively prioritizing relevant parts of prior context and identifying what information is important. We categorize these errors as shown in Table 1.

Conclusion-level Errors Conclusion-level errors indicate issues with deductive reasoning, reflecting the LLM's ability to follow premises to reach the correct conclusion. They also reveal how much the decision-making process is influenced by intermediate premises in choosing the final answer. We categorize these errors as shown in Table 2.

Conclusion from Incomplete Premises *vs.* Correct Premises We argue that a 'Wrong Conclusion from Correct Premises' is essentially a 'Wrong Conclusion from Incomplete Premises' because either the premises, though correct, are incomplete and lead to a wrong conclusion, or the LLM fails to explicitly generate a key premise. This poses a challenge for LLM-based auto-evaluators, as discussed in §3.4, which struggle to assess whether the rationale is sufficient or inadequate.

3.4 LLM-aided Automatic Evaluation

Manual analysis of reasoning chains provided a detailed categorization of errors; however, it was time-consuming and, therefore, challenging to scale for

Broad-Category	Sub-Category	Description
Wrong Conclusion	Wrong Conclusion from False Premise(s)	This error primarily occurs when the step-by-step rationale generated includes premises that are logically invalid, factually incorrect, irrelevant to solving the question posed, or a combination of these issues.
	Wrong Conclusion from Incomplete Premise(s)	This error occurs when valid and sound premises are provided but fail to fully support the reasoning, leading to the wrong conclusion. A special example is 'Wrong Conclusion from Correct Premises,' where sufficient premises still result in an incorrect conclusion.
Right Conclusion	Right Conclusion from False Premise(s)	This error occurs when the LLM reasons to the correct option while providing a wrong argument. One or multiple premises contain errors which fall under one of the three premise-level error categories and yet lead to the reasoning path choosing the correct option as its final answer.
	Right Conclusion from Incomplete Premise(s)	This error occurs when the correct final option is selected, even though the premises provided are incomplete or insufficient to fully justify that conclusion. *Our current auto-evaluator system cannot detect this, as it cannot distinguish it from a 'Correct Conclusion from Correct Premises' scenario.*
	Right Conclusion with Hallucinated Content	This error occurs when the reasoner selects the correct option but the generated content does not semantically match the provided options. For example, the LLM might output 'Option D. The suspect is *X*,' when the actual content is 'Option D. The suspect is *Y* and committed crime *Z*,' leading to a deviation due to LLM hallucination.

Table 2: Error taxonomy for the Conclusion.

the entire dataset. Thus, we develop an alternate approach to leverage LLMs to evaluate the errors in the reasoning chains akin to human evaluation. Specifically, we use GPT-4o to identify and label the errors. The details of the implementation are described in Appendix E and an example snippet of LLM-aided annotation is provided in Table 3. We develop two approaches for error evaluation:

Exact Error Label Match In this approach, we task the 'auto-evaluator' with identifying the exact error category labels which the human evaluators had labeled a particular premise/conclusion of a reasoning chain. Experiments revealed significant mislabeling between the auto-evaluator and human evaluators, with many 'Misinterpretation' errors at the premise level being labeled as 'Irrelevant Premises' or 'Factual Hallucination' (Refer Appendix D) by the auto-evaluator, and vice versa. Hence, we make changes to the auto-evaluators to include error explanations along with the labels.

Another significant challenge was the low error detection rate of factual hallucinations with the help of single-call LLM auto-evaluators. Motivated by Varshney et al. (2023); Dhuliawala et al. (2023), we develop a multi-call LLM system, consisting of two separate LLM calls, in which one LLM call creates verification questions to probe various aspects of a premise and another LLM call answers them citing the provided legal context for factuality. A premise is considered to contain factual hallucination if the answers to any of the verification questions contradicts the content of the premise directly.

Semantic Error Explanation Match As an alternative approach to the above problems, we develop a multi-analyzer system consisting of three 'single-call' and one 'multi-call' LLM-based pipeline focused on providing explanation of errors at the premise-level. A 'summarizer' LLM (Refer Appendix F) combines the individual analyses of all analyzers into a single error explanation for a premise. This enables the pipeline to detect and label multiple errors in a single premise (e.g., a premise containing both misinterpretation and factual hallucination).

To validate the effectiveness of the auto-evaluator, we sample the 30 reasoning chains from the manually evaluated set of Mistral-7B-v2-Instruct. The human evaluators then compared their error category assignments as well as explanations to those given by the auto-evaluator. The agreement score (recall %) for detecting both the presence and absence of an error in reasoning steps (through explanation match) between the manual evaluation and the auto-evaluator is $\sim$85.3% and $\sim$86%, respectively. Appendix J (Figure 17) shows the pipeline of the error detection implemented using GPT-4o.

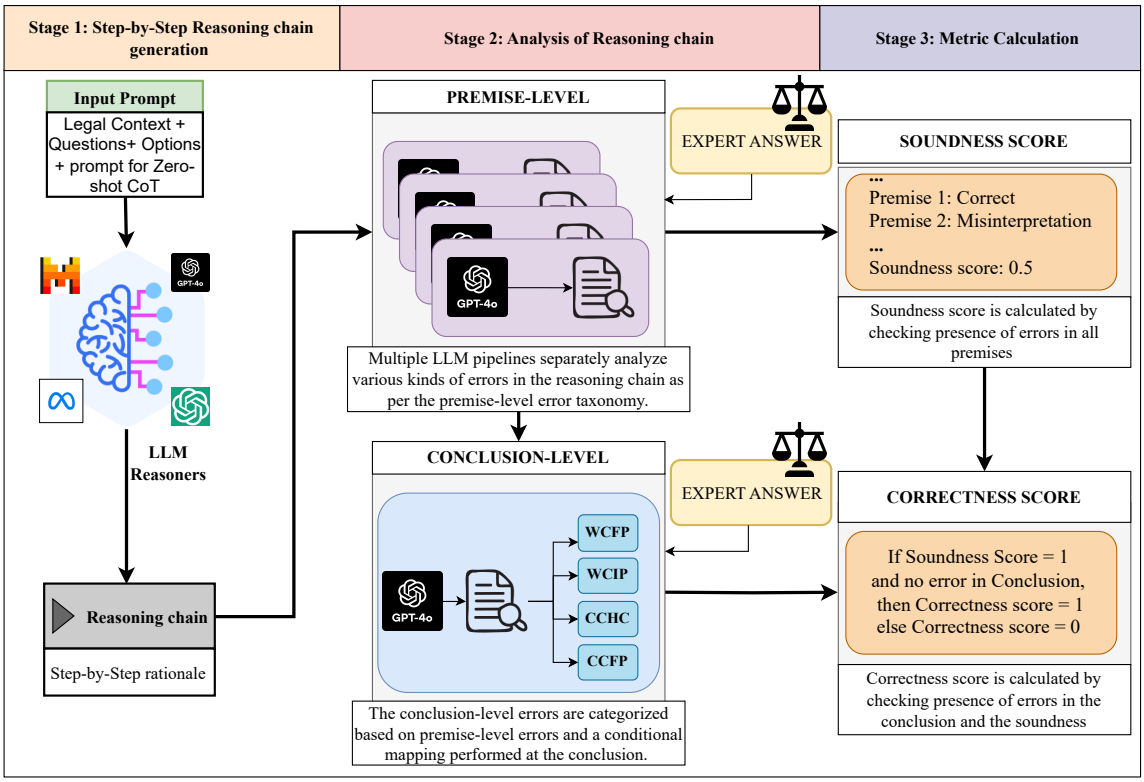

Figure 3: The overall schematic representation of the LLM-based error-detection and evaluation system and the calculation of the metrics. The reasoning chains are produced by 5 LLMs and the expert answer is referenced from the *Civ. Pro.* dataset

4 Experimental Setup

4.1 Models

We evaluate a range of closed-source LLMs including GPT-4-Turbo, GPT-4o, GPT-3.5-turbo, and open-source models Llama-3-8B-Instruct, and Mistral-7B-v2-Instruct on the '*Civ. Pro.*' dataset using the zero-shot CoT prompting setting (Wei et al., 2023; Kojima et al., 2023). To solve a given legal problem, a LLM needs to generate a set of statements $<A : s_1, s_2, ..., s_k, c>$, where A represents the legal argument/rationale put forward to solve the problem, with $s_1, s_2 ... s_k$ being the 'k' number of intermediate steps generated to reason towards the final conclusion c. NVIDIA A100 GPUs were used for conducting the inference of open-source models with a batch size of 1. An example prompt used for these experiments is provided in the Appendix J.

4.2 Metrics

Accuracy We use accuracy to demonstrate the capability of LLMs in solving legal scenarios based on their ability to predict the final answer (the conclusion). To calculate this metric, we use the LLM-generated final answers and compare them with the available expert answer (gold solution).

Soundness Score We created a step-wise soundness score metric to check the number of premises which are error-free in terms of the absence of errors as delineated by our established error taxonomy. We computed this score by taking its average across a single reasoning chain. The Soundness score (between 0 and 1) is calculated by:

$$S = \frac{\text{Number of sound premises}}{\text{Total Number of premises}}$$

Correctness Score This metric is calculated to evaluating the condition where the reasoning chain must be both error-free at both premise and conclusion level. Likewise in the soundness metric, the expert answers to extract the ground-truth options as conclusions. A reasoning chain must be both 'sound' (with soundness score being equal to 1) and arrive at the correct final option as its conclusion. The Correctness score is calculated as:

$$C = \begin{cases} 1, & \text{if (premises + conclusion) correct} \\ 0, & \text{for all other cases} \end{cases}$$

Examples of a reasoning step and final conclusion evaluated by GPT-4o based auto-evaluator
Premise 4: Marla's Intent: Marla's intent is somewhat ambiguous. She is attending a two-year program and is unsure if she wants to be a stylist. She has considered staying in Denver to work if she likes the program, or leaving the program and looking for work in Denver if she doesn't like it. **Auto-Evaluator Evaluation and Error Explanation:** Premise 4: Misinterprets Marla's intent. The premise incorrectly suggests ambiguity in Marla's intent. According to the legal context, Marla's plans are open-ended, meaning she does not have definite plans to leave Denver. The ambiguity mentioned does not align with the legal context's explanation of "indefinite intent". **Error Category Label:** "Misinterpretation"
Predicted Conclusion: Given that Marla resides in Colorado and has no definite plans to leave, she meets the domicile test's requirement of residing with the intent to remain indefinitely. Final answer: D, **Extracted Ground-Truth Conclusion:** D. is domiciled in Colorado, because she resides there with the intent to remain indefinitely. **Auto-Evaluator Evaluation and Error Explanation:** Based on the provided conditions and the information given: - The extracted conclusion's final answer is D. - The ground truth option is also D. - The content of the extracted conclusion and the ground truth option match: both state that Marla is domiciled in Colorado because she resides there with the intent to remain indefinitely. - There is an error in the premises (misinterpretation of Marla's intent). Given these points, the appropriate classification is: **CORRECT CONCLUSION FROM FALSE PREMISES** **Error Category Label:** "Correct Conclusion from False Premises"

Table 3: Example of reasoning step (premise) and conclusion evaluated by LLM-based 'Auto-evaluator' (GPT-4o). The error category labels are extracted from the detailed explanations using a LLM prompted to extract error keywords.

5 Results and Analysis

5.1 Objective Evaluation

Soundness metrics are high but correctness scores are low Table 4 shows that the majority of the premises are error-free (with the highest being GPT-4o having 78.4% of the generated premises being error-free). In contrast, Figure 12 reveals that an average of ∼96% of reasoning chains leading to conclusions from false premises have one or more misinterpretation errors in the intermediate premises. This finding, aided by empirical human analysis, suggests that much of the LLM-generated reasoning chain re-iterates existing context, while most errors occur in the smaller portion where new 'decision-making' inferences are generated. The similarity in correctness score in Mistral-7B-v2 and Llama-3-8B in contrast to the higher accuracy of Llama-3-8B could be attributed to the lesser number of steps (see Table 6 (Appendix C)) on average in the reasoning chain of Llama-3-8B when compared to Mistral-7B-v2-instruct.

Model	S (↑)	A (↑)	C (↑)
Mistral-7B	0.623	0.371	0.131
Llama-3-8B	0.493	0.451	0.137
GPT-3.5-turbo	0.607	0.417	0.217
GPT-4-turbo	0.738	0.725	0.417
GPT-4o	**0.784**	**0.737**	**0.445**

Table 4: The results for soundness, accuracy, and correctness metrics for all LLMs on the *Civ. Pro.* dataset. Here 'S' denotes the Soundness, 'A' denotes the Accuracy, and 'C' denotes the Correctness.

Accuracy *vs.* Correctness Score Table 4 and Figure 4 show a sharp decrease (an average of ∼27%) in the scores of accuracy to correctness across all LLMs. The highest fall in percentage is observed in Llama-3-8B (31.4% decrease). This is significant as it shows that while LLMs can arrive at the correct conclusion, there are a lot of cases where the reasoning chain they generate is not entirely error-free. This suggests that LLMs often rely on superficial correlations and patterns, likely learned during pre-training, to arrive at correct conclusions, rather than through genuine reasoning.

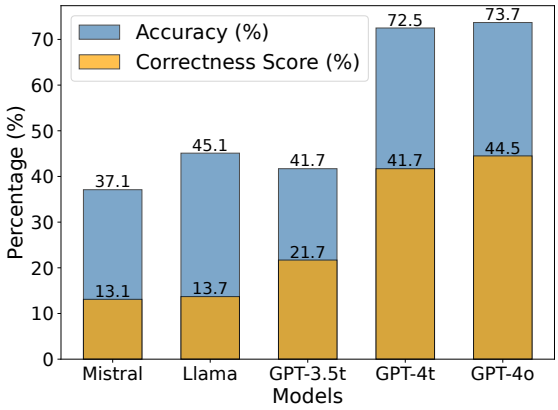

Figure 4: Performance of 5 LLMs in terms of (Accuracy metric) vs. (Correctness metric) on the *Civ. Pro.* dataset.

Larger, proprietary models 'reason' better than smaller, open-source models Figure 4 and Table 4 clearly convey that proprietary models generate more error-free reasoning steps and arrive at the correct conclusion more often than the open-source LLMs. An exception is GPT-3.5-turbo, which performs comparably to Llama-3-8B and Mistral-7B, suggesting that training data and methods might play a more significant role in enhancing reasoning than merely scaling model parameters.

Prompting techniques →	Baseline (CoT)	Plan-and-Solve	Self-Correct	Self-Discovery
Gemini-1.5-Flash				
w/o feedback	63.31	59.17	**61.54**	**64.50**
w/ feedback	**64.50**	**62.13**	59.76	63.31
Llama-3-8B-Instruct				
w/o feedback	53.71	50.29	48.58	47.42
w/ feedback	**57.14**	**52.00**	**52.57**	**49.14**

Table 5: Comparison of accuracy metric for models under different prompting strategies with and without feedback.

5.2 Reasoning Chain Evaluation

'Misinterpretations' are the dominant category of errors at premise-level Figures 5 (Appendix A) and Figure 12 (Appendix J) reveal that 'Misinterpretation' is the most dominant category of error which occurs in the reasoning chains at the premise-level. This highlights that, despite being trained on vast data, LLMs struggle to fully grasp the nuanced complexities of legal scenarios (or other contexts requiring critical analysis) in zero-shot settings.

'Wrong Conclusion from False Premises' is the dominant category of error at conclusion-level The prevalence of 'Wrong Conclusion from False Premises' (Figure 6 (Appendix B)) in conclusion-level errors results from premise-level mistakes leading to incorrect conclusions. However, in GPT-4-turbo and GPT-4o, the dominant error is 'Correct Conclusion from False Premises,' suggesting these models may be relying on patterns of similar examples from their training.

5.3 Discussion on Error-Mitigation Strategies

We carry out several experiments on the legal reasoning dataset, employing widely used prompting techniques alongside the most frequently observed errors we found through §3.3 with the aim to explore the possibility of enhancing the reasoning capabilities of both closed-source and open-source LLMs. Four prompting techniques are utilized: (1) Chain-of-Thought (Wei et al., 2022) (2) Plan-and-Solve (Wang et al., 2023) (3) Self-Correct (Zhang et al., 2024) and (4) Self-Discovery (Zhou et al., 2024). These techniques are tested with and without incorporating error definitions as feedback, following the Feedback-Learning method (Tyagi et al., 2024). Detailed descriptions of the prompting strategies can be found in Appendix G.

The error definitions are provided in three styles: generic, short, and long. The generic version uses the error definitions from the Feedback-Learning method, while the short and long versions are derived from the error taxonomy described in §3.3. All experiments are conducted in a zero-shot setting, and we evaluate each prompting technique based on the accuracy metric. We test one closed-source model, Gemini-1.5-Flash, and one open-source model, Llama-3-8B-Instruct.

As shown in Table 5, adding the error definitions as feedback showed improvement in accuracy up to 4%. For Llama-3-8B-instruct, accuracy improved across all prompting techniques, whereas for Gemini-1.5-Flash, the accuracy increased only for the Chain-of-Thought and Plan-and-Solve methods. From our observations, the decrease in accuracy for these strategies with Gemini resulted due to self-doubting (Krishna, 2023) nature of LLMs. These findings suggest that while feedback on errors provides marginal improvements in LLM performance, there is a need to develop more effective frameworks beyond prompting, such as agent-based methods, that account for these errors and enhance the model's legal reasoning capabilities.

6 Conclusion

Through our work, we assess the reasoning capabilities of LLMs by examining their performance on the '*Civ. Pro.*' dataset designed to evaluate legal reasoning. The sequential workflow of manually evaluating LLM-generated reasoning chains on data samples, creating an error taxonomy based on the commonly occurring errors, automating the error evaluation through LLM-based pipelines lead us to some interesting insights. LLMs still struggle with producing error-free rationales while reasoning about legal scenarios despite being provided with necessary legal knowledge and context. Additionally, we present our findings on implementing various prompting techniques augmented with feedback from our error taxonomy. We hope that our work lays a solid foundation for developing a framework to critically evaluate complex reasoning tasks, such as legal reasoning, and is extendable to other domains in a similar manner.

Limitations

While the *Civ. Pro.* dataset facilitates the evaluation of the legal reasoning abilities of LLMs, legal reasoning in the real world, especially the scenarios covered in this paper (law scenario judgment prediction) often inherently contains ambiguities. There are many legal scenarios in this dataset that are complex in terms of being ambiguous and tricking the reader/reasoner and would probably require legal expertise and experience to understand the nuances required to solve such scenarios. Although our study intends to capture the errors committed by natural-language based reasoners, the systematic evaluation could benefit from converting natural language to formal language through auto formalization. While our work mainly focuses on the soundness of the reasoning steps of legal reasoning chains, we would like to acknowledge that properties like consistency, coherence, completeness, and clarity are not directly measured in the current work. Future research could benefit from exploring additional automated methods, such as formal language solvers, once auto formalization is achieved. Furthermore, the current evaluation of reasoning is restricted to English, leaving room to expand this work into a multilingual context.

Ethics Statement

We obtained the necessary permissions to use the dataset provided by Bongard et al. (2022). We have utilized AI assistants, specifically Grammarly and ChatGPT, to correct grammatical errors and rephrase sentences.

References

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can gpt-3 perform statutory reasoning?

Leonard Bongard, Lena Held, and Ivan Habernal. 2022. The legal argument reasoning task in civil procedure.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021. Lexglue: A benchmark dataset for legal language understanding in english. *arXiv preprint arXiv:2110.00976*.

Steffi Chern, Ethan Chern, Graham Neubig, and Pengfei Liu. 2024. Can large language models be trusted for evaluation? scalable meta-evaluation of llms as evaluators via agent debate.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*.

Joseph W. Glannon. 2019. *The Glannon Guide to Civil Procedure: Learning Civil Procedure through Multiple-Choice Questions and Analysis*, fourth edition edition. Wolters Kluwer Law & Business, New York.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. Roscoe: A suite of metrics for scoring step-by-step reasoning.

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. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models.

Carole D. Hafner and Douglas H. Berman. 2002. The role of context in case-based legal reasoning: teleological, temporal, and procedural. *Artificial Intelligence and Law*, 10(1-3):19–64.

Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, Zhen Wang, and Zhiting Hu. 2024. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*.

Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Zhuang Li, and Adnan Trakic. 2024. Bridging law and data: Augmenting reasoning via a semi-structured dataset with irac methodology. *arXiv preprint arXiv:2406.13217*.

Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Adnan Trakic, Terry Yue Zhuo, Patrick Charles Emerton, and Genevieve Grant. 2023. Can chatgpt perform reasoning using the irac method in analyzing legal scenarios like a lawyer? *arXiv preprint arXiv:2310.14880*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Satyapriya Krishna. 2023. On the intersection of self-correction and trust in language models.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S Yu. 2023. Large language models in law: A survey. *arXiv preprint arXiv:2312.03718*.

Jack Lanchantin, Shubham Toshniwal, Jason Weston, Sainbayar Sukhbaatar, et al. 2024. Learning to reason and memorize with self-notes. *Advances in Neural Information Processing Systems*, 36.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making large language models better reasoners with step-aware verifier.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive verification of chain-of-thought reasoning.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Taiming Lu, Muhan Gao, Kuai Yu, Adam Byerly, and Daniel Khashabi. 2024. Insights into llm long-context failures: When transformers know but don't tell. *arXiv preprint arXiv:2406.14673*.

Philipp Mondorf and Barbara Plank. 2024. Beyond accuracy: Evaluating the reasoning behavior of large language models–a survey. *arXiv preprint arXiv:2404.01869*.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. Receval: Evaluating reasoning chains via correctness and informativeness.

Nemika Tyagi, Mihir Parmar, Mohith Kulkarni, Aswin RRV, Nisarg Patel, Mutsumi Nakamura, Arindam Mitra, and Chitta Baral. 2024. Step-by-step reasoning to solve grid puzzles: Where do llms falter?

Gladys Tyen, Hassan Mansoor, Victor Cărbune, Peter Chen, and Tony Mak. 2024. Llms cannot find reasoning errors, but can correct them given the error location.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation.

Vern R. Walker. 2007. Discovering the logic of legal reasoning. *Hofstra Law Review*, 35(4).

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D Goodman. 2024. Quiet-star: Language models can teach themselves to think before speaking. *arXiv preprint arXiv:2403.09629*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Jaekyeom Kim, Moontae Lee, Honglak Lee, and Lu Wang. 2024. Small language models need strong verifiers to self-correct reasoning.

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V. Le, Ed H. Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. 2024. Self-discover: Large language models self-compose reasoning structures.

A Percentage distribution of Premise-level errors

Figure 5 represents the percentage distribution of premise-level errors across the reasoning chains of all 5 LLMs. Due to the lowest number of average steps in reasoning steps, Llama-3-8B-instruct has the highest proportion of errors in the reasoning chains (∼65.4%) containing premise-level errors in the reasoning steps.

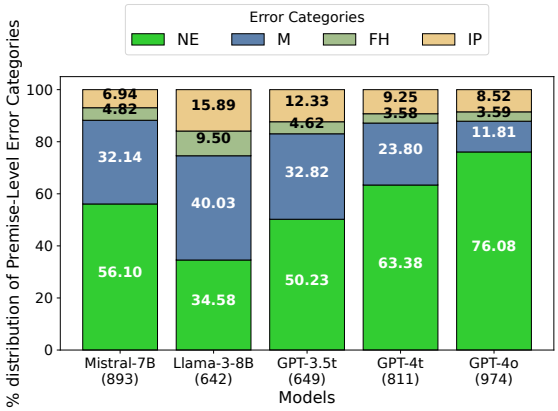

Figure 5: The percentage distribution of the premise-level error categories across the reasoning chains of all 5 LLMs. The total number of steps generated by each model is provided inside the round brackets below the model names. Here 'NE' denotes Correct Premise (No errors), 'M' denotes Premise containing a Misinterpretation, 'FH' denotes Factual Hallucination in the premise, 'IP' denotes an Irrelevant Premise.

B Percentage distribution of Conclusion-level errors

Figure 6 represents the percentage distribution of conclusion-level errors across the reasoning chains of all 5 LLMs. Mistral-7B-v2-instruct (∼86.8%) and Llama-3-8B-instruct (∼86.3%) have the highest proportion of errors in the reasoning chains containing conclusion-level errors. An interesting observation is that only Llama-3-8B instruct and GPT-4o have non-zero percentages of errors, which could possibly mean that these LLMs have been trained on very similar data to that in the '*Civ. Pro.*' dataset which could be causing these LLMs to spuriously output modified content for the correct options.

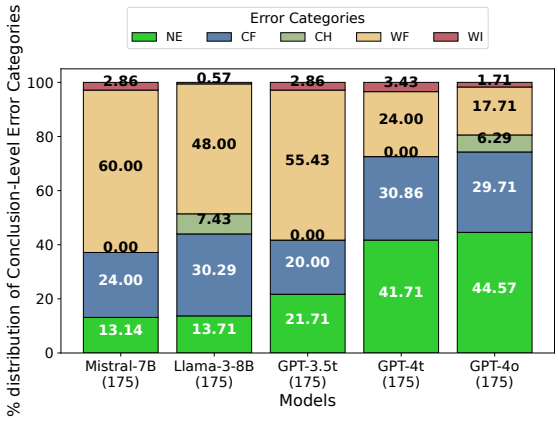

Figure 6: The percentage distribution of the conclusion-level error categories across the reasoning chains of all 5 LLMs. The total number of steps generated by each model is provided inside the round brackets below the model names. Here 'NE' denotes Correct Conclusion(CC) From Correct Premises(CP) (No errors), 'CF' denotes CC from False Premises (FP), 'CH' denotes CC with Hallucinated Content, 'WF' denotes Wrong Conclusion(WC) from FP and 'WI' denotes EC from Incomplete Premises.

C Average number of steps generated by LLMs

The average number of steps (the premise-level steps) generated by each LLM to solve a legal scenario is provided in Table 6. Llama-3-8B-instruct has the lowest average of number of steps which is a probable cause leading to lower soundness score as shown in Table 4. Overall, LLMs mostly try to complete the reasoning chain in around 4-6 steps to arrive at the conclusion. While this is good for not introducing redundancies and keeping the irrelevant premise and hallucination errors low, this could also potentially be an indicator that LLMs do not explicitly 'deliberate' enough to solve potentially harder problems.

Model	Average number of steps
Mistral-7B (893)	5.1
Llama-3-8B (642)	3.66
GPT-3.5-turbo (649)	3.70
GPT-4-turbo (811)	4.63
GPT-4o (974)	5.56

Table 6: The average number of steps(premises) generated by all LLMs in Zero-shot CoT setting. The numbers in brackets indicate the total number of steps generated by each LLM in the generated reasoning chains (excluding the final conclusion step) for the 175 sample *Civ. Pro.* dataset

D Misinterpretations vs. Factual Hallucinations vs. Irrelevant Premises

A premise is classified as containing a 'Misinterpretation' error when the LLM reasoner is making wrong inferences based on the information it generates. There is a 'Factual Hallucination' when the information generated by the LLM reasoner (by directly extracting from the provided content of the input legal context, questions, and options) is factually incorrect and can be easily verified while directly going through the input context. A key distinction between Misinterpretation and Factual Hallucination is illustrated in Figures 14 and 15. Detecting a 'Misinterpretation' requires the expert answer to account for complex legal reasoning nuances, while detecting a 'Factual Hallucination' does not depend on the expert answer for validation. An irrelevant premise is said to occur when a premise contains unnecessary or tangential information that does not contribute to reasoning toward the correct answer.

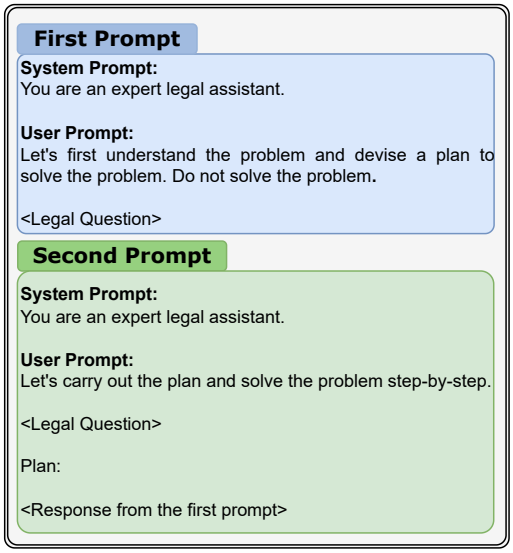

Figure 7: Prompt used to implement 'Plan-and-Solve' technique.

Extensive discussions and iterations occurred to define the taxonomy and differentiate between the three premise-level errors. However, these errors can often overlap or appear together in the same premise, especially in cases where 'Misinterpretations' are caused by 'Factual Hallucinations'. In the cases of significant overlap, both the human and the auto-evaluator were instructed multiple errors for the same premise. Detecting multiple errors can be crucial for longer premises which creates inferences on a greater number of contextual factors and nuances and hence are vulnerable to more of errors occurring in them.

E The LLM-based 'Auto-Evaluator'

At the premise level, four separate prompt-based evaluation pipelines (LLM-based error detectors) have been implemented. Three of the pipelines utilize a single call to GPT-4o, where the information provided to the LLM is broken into three parts: a. The system prompt which contains the information about the error taxa **(the knowledge base)** b. **important instructions** provided to the LLM evaluator on how to evaluate and detect errors and c. We also provide an **in-context learning example** of how the human annotation was carried out for the model to follow and replicate the annotation format.

The first pipeline (Figure 17) is designed to simply detect whether a premise contains an error, without assigning a specific label from the error taxonomy. The second pipeline detects whether a premise contains a 'Misinterpretation' error. The third pipeline does the same for detecting an 'Irrelevant Premise'. Along with this, the input prompt includes the step-by-step reasoning chain. The LLM is expected to classify the error and provide an explanation for the classification. The fourth pipeline 15, which has been designed to detect 'Factual Hallucination' errors, consists of a 'Multi-call' LLM system where primarily two calls to GPT-4o are made: 1. The first call to GPT-4o is used to create fact-verification questions about various aspects of a particular premise. 2. The second call to GPT-4o is used to answer these verification questions by referencing the legal context and content of the questions and the options. A premise is classified as containing factual hallucination(s) if the answers to its verification questions reveal a contradiction between its content and that of the provided input context. The premise-wise evaluation results of four pipelines are aggregated and summarized by a final 'summarization' call to GPT-4o which summarizes and enumerates all possible errors detected in a single premise. The aggregated and summarized results are then sent to the conclusion-level error evaluator which performs a conditional mapping procedure (as shown in Figure 16 to assign the conclusion-level errors.

First Prompt

System Prompt:
You are an expert legal assistant.

User Prompt:

<Legal Question>

Second Prompt

System Prompt:
You are an expert legal assistant.

User Prompt:
Let's verify the solution of the given question below and suggest some corrections.

<Legal Question>

Solution:

<Response from the first prompt>

Third Prompt

System Prompt:
You are an expert legal assistant.

User Prompt:
Now let's use the suggestions to solve the question again.

<Legal Question>

<Suggested corrections>

Figure 8: Prompt used to implement 'Self-Correct' technique.

F The 'Aggregator + Summarizer' LLM

A GPT-4o-based LLM (Refer Figure 17) was used to aggregate and summarize all the errors detected by the separate premise-level error detectors and summarize them for each premise separately. This had a corrective effect as the most appropriate errors (or combination of multiple errors) were summarized for each premise. An example of this corrective effect is shown in Figure 19, where both the 'Misinterpretation' and 'Factual Hallucination' auto-evaluators flag a premise with their respective labels and explanations. However, the summarizer LLM correctly identifies 'Misinterpretation' as the most accurate error classification for that premise. The LLM also filtered out unnecessary text, retaining only the premises that contained errors. This process ensured that only the premise steps flagged with errors by the auto-evaluator were forwarded as input to the conclusion-level error analyzer. By focusing on the erroneous premises, this approach streamlined the error analysis process, enabling more efficient and targeted evaluation of how these errors impact the final conclusion.

First Prompt

System Prompt:
You are an expert legal assistant.

User Prompt:
<Legal Question>

Before solving the problem select some reasoning modules

<Self-Discover reasoning modules>

Second Prompt

System Prompt:
You are an expert legal assistant.

User Prompt:
<Legal Question>

Rephrase and specify each reasoning module so it better helps solve the task:

<Response from the first prompt>

Third Prompt

System Prompt:
You are an expert legal assistant.

User Prompt:
<Legal Question>

Operationalize the reasoning modules into a step-by-step reasoning plan in JSON format

<Response from the second prompt>

Fourth Prompt

System Prompt:
You are an expert legal assistant.

User Prompt:
<Legal Question>

Follow the step-by-step reasoning plan in JSON to correctly solve the task. Fill in the values following the keys by reasoning specifically about the task. Do not simply rephrase the keys:

<JSON reasoning plan>

Figure 9: Prompt used to implement 'Self-Discovery' technique.

G Prompting-techniques for Error Mitigation

We carried out several experiments on the legal reasoning dataset, employing widely used prompting techniques alongside the most frequently observed errors we found through §3.3 with the aim to explore the possibility of enhancing the reasoning capabilities of both closed-source and open-source LLMs. Four prompting techniques were utilized: (1) Chain-of-Thought (Wei et al., 2022) (2) Plan-and-Solve (Wang et al., 2023) (3) Self-Correct (Zhang et al., 2024), and (4) Self-Discovery (Zhou et al., 2024). We used the zero-shot CoT (an example is shown in Figure 19) method as the baseline method in which the LLM is prompted to provide the final answer along with step-by-step reasoning.

11-step Annotation Guideline

1. Go through the provided annotation example thoroughly.
2. In the step-by-step reasoning chain, all the intermediate steps (from the first step to the penultimate step) are termed as **'premises'** and final step/statement which concludes the chain by choosing an option is termed as the **'conclusion'**.
3. Take your time to refer to the error categories described in the Error taxonomy provided to you.
4. The errors in the premises are **'premise-level'** errors and the error in the conclusion is the **'conclusion-level'** error.
5. If a premise contains multiple errors from the error-taxonomy, then label it with all the errors possible.
6. Refer to the expert answers provided by the authors of the questions to double-check your error classifications.
7. Please remember to follow this order for error classification: First, analyze and annotate errors at the premise level, and then proceed to annotate errors at the conclusion level..
8. Based on the presence of errors in the premises, classify the error in the conclusion into one of the sub-categories of conclusion-level errors present in the error taxonomy.
9. Additionally, compare the content of the conclusion (the content of the option generated by the LLM reasoner while choosing them from one of the provided options as answers to the question) to check for 'Right Conclusion with Hallucinated Content' error.
10. Double-check if your error classification is the most appropriate one.
11. While the task is focused on identifying errors in the reasoning-chain rationale rather than solving the legal question itself, you are encouraged to independently attempt referring to the legal context and solving the question to enhance your understanding of the legal scenario.

Figure 10: The 11-step guideline provided to the annotators for conducting manual evaluations of the LLM-generated rationale

Plan-and-Solve prompts the LLM first to generate a plan to solve the problem without solving it and after that the LLM carries out the self-suggested plan to get the final answer as shown in Figure 7. Self-Correct uses self-verification and self-refining to improve the reasoning ability of the LLMs as shown in Figure 8. Self-Discover as shown in Figure 9 utilizes self-discover reasoning modules to create an explicit reasoning structure to follow to solve the problem. For running the above prompting strategies with error taxonomy as the feedback, we include a detailed description of the error taxonomy in the system prompt. For instance, Plan-and-Solve with error feedbacks is shown in Figure 11.

H Guidelines for Annotation

A set of 11 annotation guidelines was developed and provided to annotators for manual evaluation and annotation. Four annotators, two students from undergraduate and graduate level each respectively, participated in this process, and cross-evaluation was conducted at the final stage to resolve any discrepancies. These guidelines also served as the basis for creating system prompts used in the LLM-based auto-evaluator, as detailed in Section E of the Appendix.

First Prompt

System Prompt:
You are an excellent legal assistant. However, we have analyzed how you solve the legal reasoning question before and we found that you make the following mistakes more often while reasoning:

1. Hallucination....
2. Misinterpretation....
3. Irrelevant Premise....

Now, you understand how to identify and remove such errors. Using all the above knowledge, please solve the question step-by-step.

User Prompt:
Let's first understand the problem and devise a plan to solve the problem. Do not solve the problem.

<Legal Question>

Second Prompt

System Prompt:
You are an expert legal assistant.

User Prompt:
Let's carry out the plan and solve the problem step-by-step.

<Legal Question>

Plan:

<Response from the first prompt>

Figure 11: Prompt used to implement 'Plan-and-Solve with error feedback' technique. Error feedback can be added to other prompting strategies in the same way.

I Human Annotation Examples

Mistral-7B-v2-instruct was selected for human evaluation due to its unique position as the earliest and smallest parametric model among the LLMs tested. Its smaller size increases the likelihood of it producing a wider range of reasoning and contextual errors, making it an ideal candidate for error analysis. By starting with a model that has fewer parameters and is more prone to subtle reasoning gaps, we can thoroughly evaluate and better understand the types of errors that may occur. This approach maximizes the chances of capturing diverse error types that could be missed in larger, more sophisticated models, which tend to exhibit fewer surface-level mistakes. Table 7-11 contain few examples of human analysis of complexity of legal reasoning and error annotations performed for the reasoning chains. The text written in blue font in the tables represents the human analyses and annotations performed.

J Miscelleneous Figures and Tables

The subsequent pages of the Appendix contain figures and tables which could not be accommodated in the previous sections. This section contains Figures 12 to 19 and Tables 7 to 11.

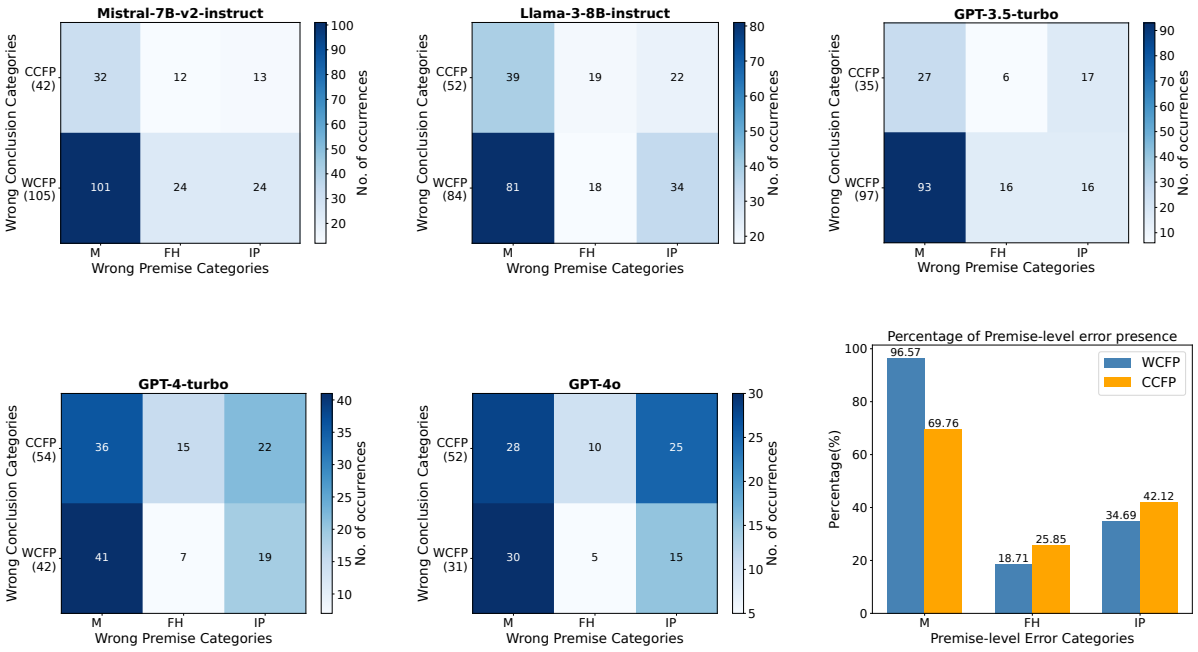

Figure 12: The first five sub-figures in the preceding section display the error distribution for premise-level errors in instances where the conclusion contains an error, across five different LLMs. The dimension in the x-axis represents the categories of errors at the premise level. The dimension in the y-axis represents the errors at the conclusion level. Here 'M' denotes 'Misinterpretation', 'FH' denotes Factual Hallucination, 'IP' denotes Irrelevant Premise, 'WCFP' denotes Wrong Conclusion from False Premise(s), and 'CCFP' denotes Correct Conclusion from False Premises. Here the counts in the heatmap represent the presence of one (or more) premise-level error(s) (in the x-axis) in a reasoning chain containing the specified conclusion-level error (in the y-axis)

System-Prompt

You are an expert in identifying errors in provided reasoning chains. Specifically, these reasoning chains are obtained from a legal MCQA task. From the reasoning chain, you have to classify each reasoning step with the presence/absence of a misinterpretation error in the step. You also have to explain why a step is classified as such. If you do not find any misinterpretation error in reasoning step, assign "NO MISINTERPRETATION" label to it. To understand some errors which can occur, please refer to provided knowledge base. While finding errors in provided reasoning chains, you can refer to gold-standard reasoning chain for better understanding.

Instruction

Trigger:
The user submits a legal context, question and options and reasoning chains 1 and 2. Reasoning Chain 1 is the gold-standard answer. Reasoning Chain 2 is the reasoning chain provided by a LLM to solve the question.

Instructions:
Take your time to reiterate the clues given in the legal problem. Follow these IMPORTANT POINTS given below:

1. Keep in mind that in the reasoning chain, only final statement with the final answer is the conclusion. All preceding statements are premises.

2. You have to find the errors of misinterpretations in the premises, not the conclusion.

3. Assign the most appropriate type of misinterpretation in reasoning chain 2. If there is no fitting type, assign the **NO MISINTERPRETATION** category.

Knowledge-Base

This Knowledge Base contains the types of misinterpretations predominantly committed by Large Language Model(LLM) based Reasoners while generating reasoning chains.

Misinterpretation is the error where the LLM misinterprets some part or entirety of the legal context or the question provided to it.
This can occur due to ambiguities, limitations in the model's training, or the complex nature of language. Misinterpretation can affect the reliability of the model's output, leading to incorrect information and reasoning.

Misinterpretation errors can occur in the following ways:
1. Failing to consider some important aspects of the given context, leading to a misunderstanding.
Example: Misunderstanding a sarcastic comment as a serious statement because the context of sarcasm was missed.

2. Misinterpreting the logical relationship between ideas, such as cause and effect, or failing to follow an argument's reasoning correctly.
Example: Assuming that correlation implies causation.

3. Misunderstanding the meaning of a word or phrase.
Example: Confusing homophones like "bare" and "bear."

4. Drawing incorrect inferences from the prior-generated text by LLMs.
Example: Inferring that someone is angry based on their terse email, when they were simply in a hurry.

5. Misunderstanding, specifically, numerical data, statistical information or mathematical concepts.
Example: Misinterpreting percentages or probability, like thinking that a 30% chance of rain means it will rain 30% of the time.

6. Misunderstanding the timing or sequence of events.
Example: Misinterpreting past events as future intentions or vice versa.

7. Misunderstanding specialized terminology used within a particular field.
Example: Misinterpreting medical jargon without the proper context or knowledge.

Structure of Response and Annotation Example

Suppose if there are 2 premises and 1 conclusion in reasoning chain 2, structure your response in the following way:
Premise 1: <Content of premise 1>
Error Category: [MISINTERPRETATION/NO MISINTERPRETATION]
Explanation: <Justify your error category>

Premise 2: <Content of premise 2>
Error Category: [MISINTERPRETATION/NO MISINTERPRETATION]
Explanation: <Justify your error category>

Conclusion: <Content of Conclusion>
Error Category: [Conclusion is not Checked for Misinterpretations]
Explanation: <Conclusion is not Checked for Misinterpretations>

Figure 13: Prompt structure for LLM evaluator 2 from Figure 17. Appropriate Instructions and Knowledge Base for detecting 'Misinterpretation' errors are provided in the prompt. Similar prompt structures and appropriate knowledge bases have been provided to all other LLM-based evaluators.

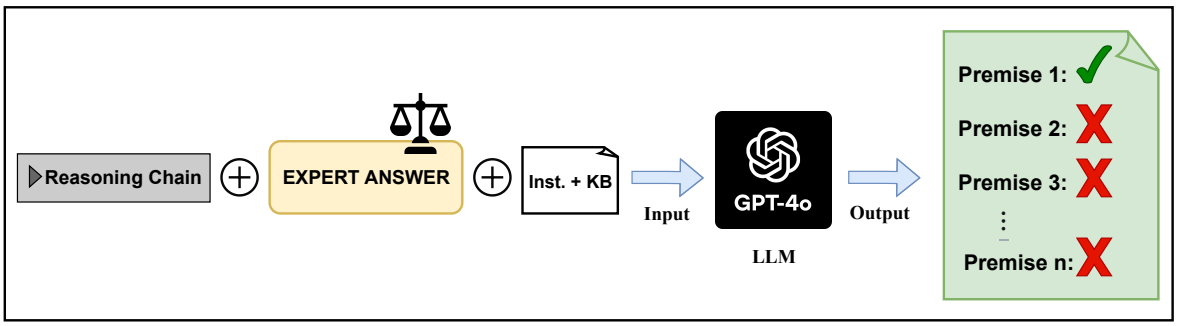

Figure 14: Schematic representation of the 'Single-Call' LLM-based auto-evaluator for premise-level error detection. 'Inst + KB' includes instructions for error analysis and the knowledge base with error definitions, based on the proposed error taxonomy. The green check-mark indicates the absence of errors and the red cross-mark represents the presence of an error in a premise.

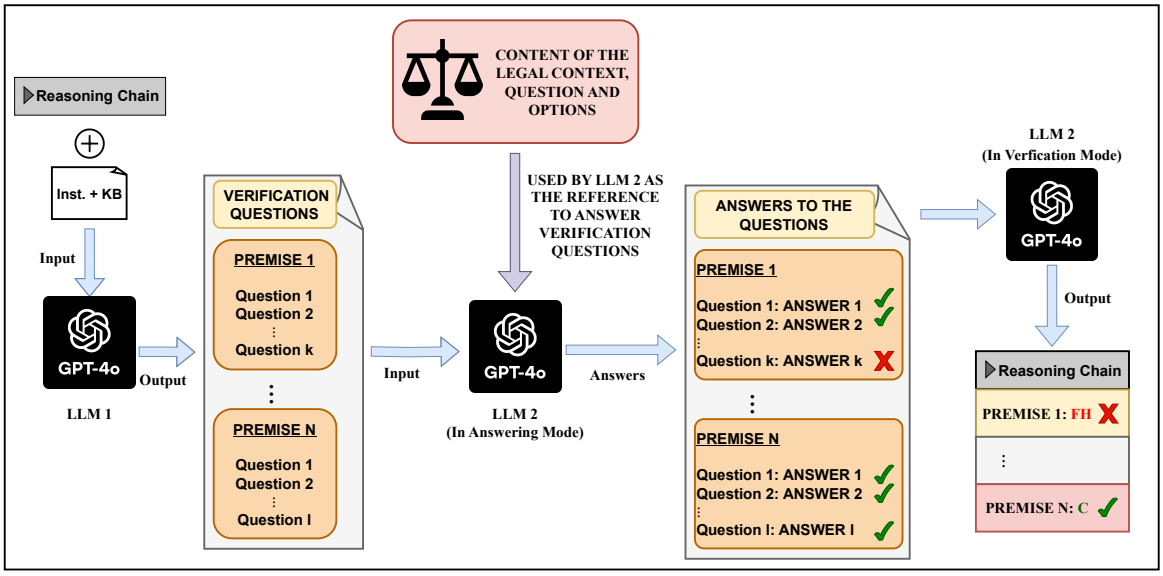

Figure 15: Schematic representation of the 'Multi-Call' LLM-based auto-evaluator for 'Factual-Hallucination' error detection. 'Inst + KB' includes instructions for error analysis and the knowledge base with error definitions, based on the proposed error taxonomy. The green check-mark indicates the verification question being correctly answered and the red cross-mark represents the presence of a contradiction in the content of the premise with the provided context.

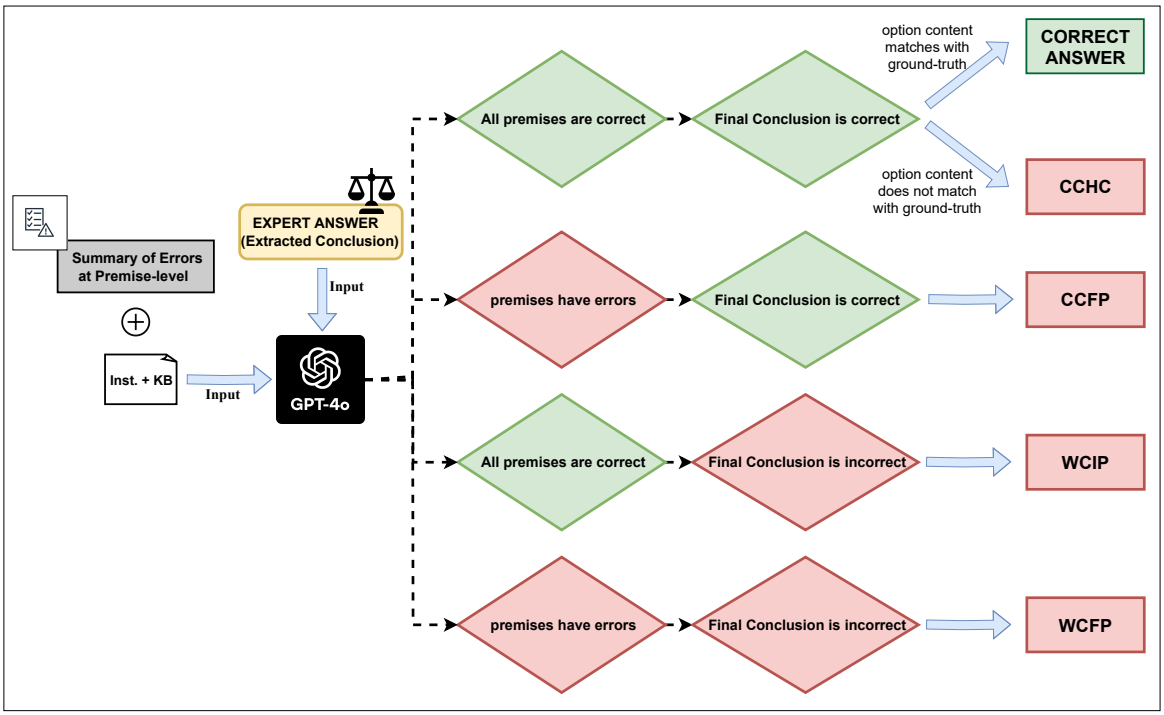

Figure 16: Schematic representation of the LLM-based auto-evaluator system for error detection at conclusion-level. The GPT-4o LLM represents the 'LLM evaluator 5' in Figure 17. The dotted lines represent the conditional paths of which only one will be true and lead to respectively conclusion-level error being labelled to the conclusion, Here, 'CCHC' represents 'Correct conclusion with Hallucinated Content, 'CCFP' represents 'Correct Conclusion from False Premises', 'WCIP' represents 'Wrong Conclusion from Incomplete Premises' and 'WCFP' represents 'Wrong Conclusion from False Premises'.

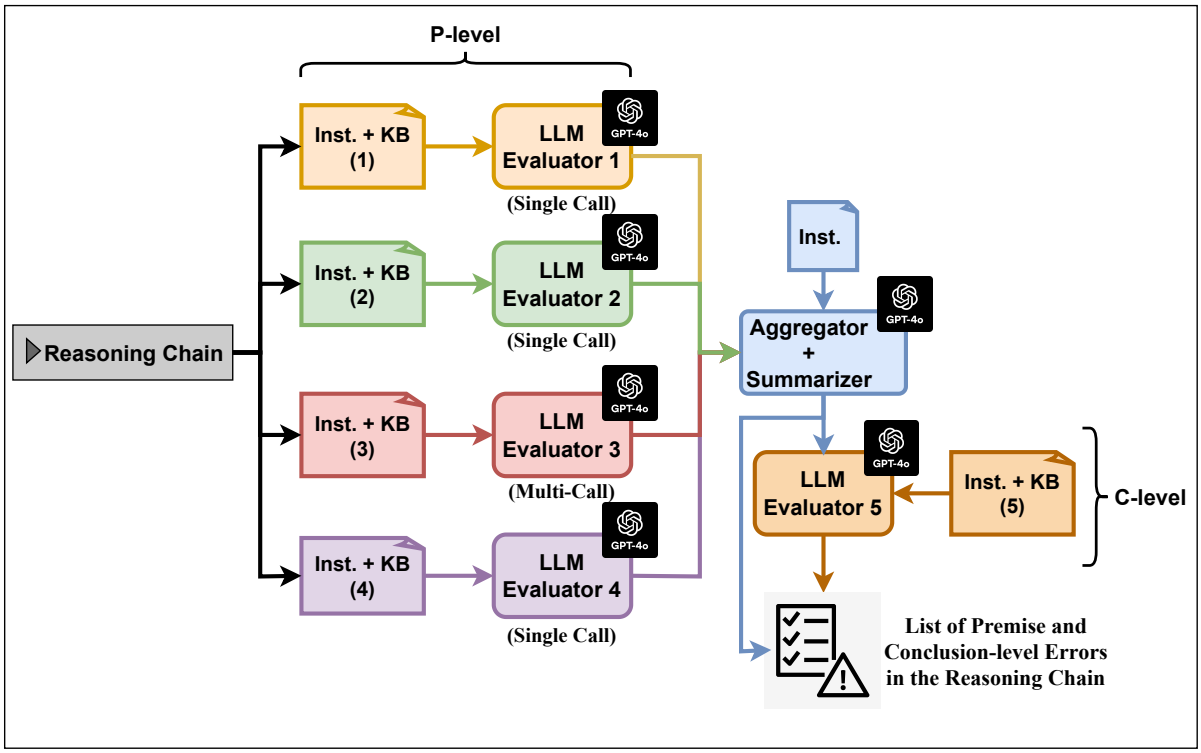

Figure 17: Schematic representation of the LLM-based auto-evaluator system for error detection. 'Inst + KB' with their respective numbers in brackets includes instructions for error analysis and the knowledge base with error definitions, based on the proposed error taxonomy. An example prompt structure (for LLM evaluator 2) has been shown in Figure 10. Here, 'P-level' denotes the Premise-level, and 'C-level' denotes the Conclusion-level. The schematics of the 'Single-Call' and 'Multi-call' GPT-4o LLM-based evaluators are represented in Figure 14 and 15 respectively.

Legal Context and Question
The Due Process Clause of the Fourteenth Amendment prohibits a state from depriving a person of life, liberty, or property without due process of law. When a court enters a civil judgment against a person, it begins the process of taking the person's property (usually in the form of money). Consequently, the due process clause requires the court to use a fair procedure in entering judgment. Certainly, one component of a fair procedure is to tell the defendant that the court is going to adjudicate her rights. Hence, the Due Process Clause requires a court to use a constitutionally adequate means of notifying the defendant that a lawsuit has been commenced against her. Typically, it is the plaintiff who does the legwork of serving process on the defendant. The statutes or court rules in every state contain detailed provisions governing how this notice of a lawsuit is provided to the defendant. Court rules may authorize various means of serving process. The most obvious is to deliver the initiating papers in the case to the defendant in person, called ''personal service of process.'' Other methods are often authorized as well, though they are less certain to actually inform the defendant about the suit. The service rules may provide, for example, that the papers may be left with someone at the defendant's home or place of business or slipped under the door. Some authorize service to the defendant by certified or ordinary mail. In some circumstances, court rules or statutes may authorize service by publication, by printing a notice of the suit in the legal notices in the local newspaper for several weeks. Such publication notice is sometimes referred to as ''constructive service,'' as opposed to actual delivery of the papers to the defendant. Which of these methods is constitutionally adequate under the Due Process Clause? Justice Jackson's wise opinion in Mullane v. Central Hanover Bank & Trust Co., 339 U.S. 306 (1950) does not provide a mechanical answer to that question, but instead establishes broad standards for constitutionally adequate notice. Mullane involved a device called a common trust fund, which allowed a bank to pool a group of small trusts into one common fund, which could be managed more efficiently. Income (or losses) from the trust would then be shared by each smaller trust in proportion to their contributions to the fund. Periodically, the trustee of the fund would file accounts with the court. If the court approved the accounts, its judgment would bar investors from suing the bank for mismanagement. Since this judicial accounting could deprive trust beneficiaries of property—the right to sue the bank for mismanagement—due process required adequate notice to them of the proceeding. The problem in Mullane was that it was difficult to ascertain who all the beneficiaries were. Beneficiaries currently entitled to receive income from the trust were known, and the bank had their addresses in its files. But most trusts have contingent beneficiaries, such as the children or heirs of a beneficiary if the primary beneficiary dies. The bank might not know who these beneficiaries were, and they could change periodically. Perhaps, with a large investment of time by lawyers and investigators, the bank could identify most them, but doing so would eat up the profits of the common trust fund. Many were contingent beneficiaries; that is, they had no current right to receive income from the fund but might become vested beneficiaries later under the terms of the individual trusts. So who did the bank have to tell about the suit and how? Mullane held that ''an elementary and fundamental requirement of due process in any proceeding which is to be accorded finality is notice reasonably calculated, under all the circumstances, to apprise interested parties of the pendency of the action and afford them an opportunity to present their objections.'' 339 U.S. at 314. What is reasonable in one circumstance would not necessarily satisfy due process in another. On the facts of Mullane, several factors reduced the risk of error if notice failed to reach all beneficiaries. First, the statute provided for the appointment of a guardian in the settlement proceeding to represent the interests of all beneficiaries. Second, even if all beneficiaries did not receive notice of the proceeding, many would, and would likely share and represent the interests of all beneficiaries in the settlement of the accounts. Consequently, the Court in Mullane did not require that each individual with an interest in the fund be given personal notice of the action. The question below probes what it did require. Question: In Mullane v. Central Hanover Bank & Trust Co., the Supreme Court held that Options: A. every person whose interests may be affected by a judicial proceeding is entitled to at least mail notice of the proceeding. B. every person whose name and address could be ascertained through reasonable investigation must be given individual notice of the proceeding. C. any person whose interests might be affected by the proceeding must be given notice by in-hand service of process. D. due to the large number of persons whose interests might be affected by the proceedings, notice by publication was sufficient on the facts of the case. E. None of the above is true.

(a) Part 1: Contains the Legal Context, Questions and Options.

Figure 18: A Legal scenario which showcases the complexity of legal reasoning involved. This is one of the cases where all the LLMs under evaluation provided wrong options as answer.

Expert Answer	
A takes the position that everyone who might be affected by the settlement of the accounts must receive individual notice. If that were true, it would be impossible to settle the accounts of the trust. Even some beneficiaries with a current right to income may be impossible to find; if due process required individual notice to them it couldn't be done. Other people have contingent interests: They might get income in the future, for example, as the heir of a current beneficiary. The bank may not know who these contingent beneficiaries are, though it might be able, with research, to find some of them. Justice Jackson viewed due process as a flexible concept that involves a balance of the importance of the right involved, the risk of an erroneous decision, the cost of providing individual notice, and the likelihood that the absentee's interest would be protected by other parties. He rejected the view that everyone with some interest in the action must be individually notified for it to proceed. B would impose a lesser burden: to notify the persons whose interests could be ascertained with reasonable investigation. This could be done, but it would be costly, involving legal review of trusts and factual research on the whereabouts of all identified beneficiaries. This would more effectively reach affected persons, but the cost of doing so would likely impose such a severe burden on the plan as to "dissipate its advantages." 339 U.S. at 318. Under the circumstances—including the presence of the guardian appointed to represent the interests of beneficiaries and the actual notice to many of them—the Court held that due process did not require that. Consequently, a beneficiary might never learn of the proceeding, have her right to object to the accounts foreclosed by it, and have no constitutional complaint. C is also a loser. Mullane held that notice by mail is sufficiently likely to actually inform a person of the proceeding to be constitutionally sufficient under most circumstances. In fact, Mullane did not hold that anyone had to be notified of the suit by personal service of process—that is, by in-hand delivery of the papers. And D fails as well. The Court required at least mail notice of the action to those beneficiaries whose names and addresses were known to the bank, since they could easily and inexpensively be informed of the action. So E is right. None of the prior statements is quite true.	
GPT-4o Answer	**Challenges**
Explanation: 1. The Supreme Court in Mullane v. Central Hanover Bank & Trust Co. established that due process requires notice reasonably calculated, under all the circumstances, to apprise interested parties of the pendency of the action and afford them an opportunity to present their objections. 2. The Court did not mandate a specific method of notice for every situation but emphasized that what is reasonable in one circumstance may not be in another. 3. The Court recognized that personal notice to every individual might not be feasible, especially when the identities and addresses of all interested parties are not easily ascertainable. 4. The Court held that for those beneficiaries whose names and addresses were known or could be ascertained through reasonable investigation, individual notice was required. 5. For those whose identities or addresses could not be reasonably determined, notice by publication was deemed sufficient under the circumstances of the case. Final answer: B	• **Overlooking 'None of the Above':** It's easy to default to the option that seems closest to the Court's holding without considering that none of the options may be entirely accurate. • **Close Similarities Between Options:** The answer choices are subtly different, making it challenging to distinguish the precise legal nuances each one presents. • **Complex Legal Standards:** The principles established in *Mullane* involve nuanced interpretations of due process and notice requirements, which can be difficult to encapsulate in a single, clear-cut answer. • **Misleading Phrases:** Terms like "reasonable investigation" or "in-hand service of process" can be misleading if not directly aligned with the Court's language, leading to incorrect assumptions. • **Assumptions Based on General Knowledge:** Relying on a general understanding of due process requirements might lead one to select an answer that doesn't precisely match the specifics of the case. • **Interpretation of 'Individual Notice':** The concept of who must receive individual notice and under what circumstances can be complex, potentially causing confusion when selecting the correct answer. • **Ambiguity in the Court's Holding:** The Court's decision balances various factors, and without a mechanical rule, it's challenging to determine which answer fully aligns with the holding.

(b) Part 2: Contains the Expert Answer, LLM step-by-step response and Reasoning challenges.

Figure 18: Continuation from the previous sub-part of Figure 18. This sub-part contains the expert-answer to a legal scenario, a wrong response provided by GPT-4o and the challenges faced by a legal reasoner to solve this scenario

Prompt to Mistral-7B-v2-instruct	Mistral-7B-v2-instruct Response
Task: You are a helpful legal assistant. Choose the correct option by performing legal reasoning while strictly adhering to the legal context below. **Legal Context:** Under federal diversity jurisdiction a citizen of one state may sue a citizen..**(truncated)**..assume that the court applies the reside-with-intent-to-remain-indefinitely test, and consider where Marla has established a residence with the requisite "indefinite" intent. **Question:** Marla, who grew up in Montana, moved to Colorado after high school to enter a two-year program for hair stylists at the Denver Beauty School. She wasn't sure if she really wanted to be a stylist, but she was anxious to get away from home, and her parents agreed to foot the bill, so off she went. She figured she'd stay if she liked it, and get a job as a stylist afterwards, in Denver or elsewhere in the West (including Montana). Or she would leave the program if she didn't like it and look for work, hopefully in Denver. She took an apartment on a six-month lease. After moving to Denver, Marla **Options**: A. remains domiciled in Montana, because the program is only for two years. B. remains domiciled in Montana, because she may return there to work as a stylist. C. remains domiciled in Montana, because she was domiciled there before she left for Denver. D. is domiciled in Colorado, because she resides there with the intent to remain indefinitely. **Response Structure:** While answering make sure to use the following format: [Explanation of your legal reasoning step by step as numbered points] [Final Answer with the correct option]	[1] Marla's current domicile is determined by the state where she resides with the intent to remain indefinitely. [2] The fact that Marla moved from Montana to Colorado for a two-year program does not automatically mean she is still domiciled in Montana. [3] The intent-to-remain-indefinitely prong is the key factor in determining domicile. [4] Marla's statement about possibly returning to Montana to work as a stylist after the program does not necessarily indicate she intends to remain indefinitely in Montana. [5] Marla's current residence in Colorado, along with her intent to remain there for the duration of her program, could potentially establish Colorado as her domicile. [6] Based on the legal context and the question, the correct answer is: D. is domiciled in Colorado, because she resides there with the intent to remain indefinitely. [Final Answer] D. is domiciled in Colorado, because she resides there with the intent to remain indefinitely. **Expert Answer with the Ground Truth Option** Don't be fooled by A. Although Marla's program is for two years...**(truncated)**... However, if she's in Denver without definite plans to leave, she establishes a new domicile there. The facts suggest that her stay is open-ended, so she has established a Colorado domicile. D takes the prize.

LLM-Based Auto-Evaluator System Results (PREMISE-LEVEL)

VERIFICATION QUESTIONS GENERATION FOR PREMISES:
Premise 1: Marla's current domicile is determined by the state where she resides with the intent to remain indefinitely.
1. Is Marla's current domicile determined by the state where she resides?
2. Is the intent to remain indefinitely a factor in determining domicile?

Premise 2: The fact that Marla moved from Montana to Colorado for a two-year program does not automatically mean she is still domiciled in Montana.
1. Did Marla move from Montana to Colorado for a two-year program?
2. Does moving to Colorado for a two-year program automatically mean Marla is still domiciled in Montana?

Premise 3: The intent-to-remain-indefinitely prong is the key factor in determining domicile.
1. Is the intent-to-remain-indefinitely prong a key factor in determining domicile?
2. Is there a prong called "intent-to-remain-indefinitely" in legal terms?

Premise 4: Marla's statement about possibly returning to Montana to work as a stylist after the program does not necessarily indicate she intends to remain indefinitely in Montana.
1. Did Marla make a statement about possibly returning to Montana to work as a stylist after the program?
2. Does Marla's statement about possibly returning to Montana indicate she intends to remain indefinitely in Montana?

Premise 5: Marla's current residence in Colorado, along with her intent to remain there for the duration of her program, could potentially establish Colorado as her domicile.
1. Is Marla's current residence in Colorado?
2. Does Marla intend to remain in Colorado for the duration of her program?
3. Could Marla's residence and intent to remain in Colorado establish Colorado as her domicile?

Conclusion: Based on the legal context and the question, the correct answer is: D. is domiciled in Colorado, because she resides there with the intent to remain indefinitely.
[No Validation done for the Conclusion]

ERROR-DETECTOR 1 (ERROR CHECKER WITHOUT LABELLING):
Summary of Errors:
1. Premise 5:
- Error Description: Misinterpretation of the legal context regarding domicile.
- Explanation: The premise incorrectly equates Marla's intent to remain in Colorado for the duration of her program with an intent to remain indefinitely. The legal requirement for establishing domicile is an open-ended intent to stay, not just for a fixed period like the duration of a program.

ERROR-DETECTOR 2 (MISINTERPRETATION ERRORS):
Summary of Errors:
1. Premise 5:
- Error Label: **MISINTERPRETATION**
- Explanation: This premise misinterprets the legal context. The intent to remain for the duration of a program (two years) does not satisfy the requirement of intent to remain indefinitely. The legal context specifies that domicile requires an open-ended intent to stay, not limited by a definite period or event.

ERROR-DETECTOR 3 (FACTUAL HALLUCINATION ERRORS):
Summary of Errors:
Premise 5:
- Error Label: **FACTUAL HALLUCINATION**
- Explanation: Incorrectly concludes that Marla's residence and intent to remain in Colorado for the duration of her program could establish Colorado as her domicile. The evidence indicates that a temporary stay for a specific purpose (like a two-year program) does not establish domicile if there is no intent to remain indefinitely.

ERROR-DETECTOR 4 (RELEVANCE-CHECKING):
Summary of Errors:
There are no errors present in the premises or conclusions of either Reasoning Chain 1 or Reasoning Chain 2. All premises are relevant and directly address the legal standard for determining domicile, which is pertinent to answering the question.

LLM-Based Auto-Evaluator System Results (AGGREGATION AND SUMMARIZATION)

ERROR AGGREGATION AT PREMISE LEVEL:
Combined Summary of Errors:
1. Premise 5:
- Error Description: Misinterpretation of the legal context regarding domicile.
- Explanation: The premise incorrectly equates Marla's intent to remain in Colorado for the duration of her program with an intent to remain indefinitely. The legal requirement for establishing domicile is an open-ended intent to stay, not just for a fixed period like the duration of a program. The evidence indicates that a temporary stay for a specific purpose (like a two-year program) does not establish domicile if there is no intent to remain indefinitely.

LLM-Based Auto-Evaluator System Results (CONCLUSION-LEVEL)

FINAL OPTION CHOSEN BY THE EXPERT:
D. is domiciled in Colorado, because she resides there with the intent to remain indefinitely.

CONCLUSION EXTRACTED FROM THE REASONING CHAIN:
Based on the legal context and the question, the correct answer is: D. is domiciled in Colorado, because she resides there with the intent to remain indefinitely.

ERROR AT CONCLUSION LEVEL:
Based on the provided conditions and the summary of errors, the appropriate classification for the extracted conclusion is:

CORRECT CONCLUSION FROM FALSE PREMISES

Here's the reasoning:
- There is an error in Premise 5, which misinterprets the legal context regarding domicile.
- The options in the extracted conclusion and the ground-truth option match.

Therefore, the correct classification is "**CORRECT CONCLUSION FROM FALSE PREMISES.**"

Figure 19: Example LLM Auto-Evaluator Response: Top: Zero-shot CoT prompt and LLM response. Bottom: Outputs from the LLM-based auto-evaluator.

Legal Context: Under federal diversity jurisdiction a citizen of one state may sue a citizen of another in federal court, even though her claim arises under state law if she has a colorable claim for more than $75,000. The state citizenship of a person—as opposed to that of a corporation—is determined by her domicile, that is, the most recent state where she has (1) resided with (2) the intent to remain indefinitely. The "residence" requirement is easily satisfied. Staying overnight in a hotel or a tent will establish "residence" in a state. It is the intent-to-remain-indefinitely prong that gives students problems. A person intends to remain indefinitely in a state if she is residing in the state on an open-ended basis, without the intent to leave at a definite time or on the occurrence of a definite event. You don't have to swear allegiance forever to a state to acquire domicile there; you only need to reside there "indefinitely," that is, on an open-ended basis. If a party is living in a state without definite plans to leave, the domicile test regards that state as her "home." She is there, not as a visitor, but as a citizen. She is, psychologically speaking, at home there, rather than passing through. She may choose to move on, as we all may, but at the moment she has no plans to do so. Remember that, until the two prongs coincide in a new state, your old domicile continues, whether you plan to return to that state or not. If Acari, from Hawaii, leaves for a one-year job acting in a play in California, planning to go to New York afterwards, he remains domiciled in Hawaii, even if he swears that he will never return to Hawaii. He hasn't acquired a domicile in California, because he doesn't plan to stay there indefinitely. He hasn't acquired one in New York either, since he doesn't reside there yet. Domicile doctrine abhors a vacuum, so it holds that Acari keeps his Hawaii domicile until the two prerequisites come together in another state. In analyzing the question below, assume that the court applies the reside-with-intent-to-remain-indefinitely test, and consider where Marla has established a residence with the requisite "indefinite" intent.

Question: Marla, who grew up in Montana, moved to Colorado after high school to enter a two-year program for hair stylists at the Denver Beauty School. She wasn't sure if she really wanted to be a stylist, but she was anxious to get away from home, and her parents agreed to foot the bill, so off she went. She figured she'd stay if she liked it, and get a job as a stylist afterwards, in Denver or elsewhere in the West (including Montana). Or she would leave the program if she didn't like it and look for work, hopefully in Denver. She took an apartment on a six-month lease. After moving to Denver, Marla
Options:
A. remains domiciled in Montana, because the program is only for two years.
B. remains domiciled in Montana, because she may return there to work as a stylist.
C. remains domiciled in Montana, because she was domiciled there before she left for Denver.
D. is domiciled in Colorado, because she resides there with the intent to remain indefinitely.

Human Annotator Analysis of LLM generated Reasoning Chain:
[1] The question asks about Marla's domicile based on her current residence in Denver, Colorado, and her intentions regarding her stay there. (Correct)

[2] According to the legal context, a person's domicile is determined by the state where they reside with the intent to remain indefinitely.(Correct)

[3] Marla moved to Denver, Colorado, to attend a two-year hair stylist program. She was unsure about her career choice and had plans to leave the program if she didn't like it. She also mentioned that she might look for work in Denver or other western states, including Montana.(Correct)

[4] Marla's lease in Denver is for six months, which indicates that she does not have an open-ended intention to remain in Colorado indefinitely.
(Misinterpretation) - (Lease in Denver is for six months (A) -> indicates that she does not have an open-ended intention to remain in Colorado indefinitely (B)) - Here B can not be true just because A is true, there are other factors which need to be true for B to be actually true. (A alone is not sufficient to prove that B is true.) (Incomplete/Insufficient cause for an effect)

[5] The fact that Marla was domiciled in Montana before moving to Denver does not automatically make her domiciled in Montana again. Domicile is determined by the present intent to remain indefinitely in a state, not by past domicile. (Correct)

[6] Based on the information provided, Marla does not meet the requirement of residing in Colorado with the intent to remain indefinitely. (Wrong Conclusion from False Premise) - Wrong Conclusion as a consequence of [4]; because of point [4], the LLM seems to have drawn a wrong conclusion, Error Propagation from [4] to [6])

[Final Answer] A. remains domiciled in Montana.

Annotator Remarks:
The LLM just disregarded point [5] while making the final conclusion in point [6]/ didn't utilize the observations made in point [5] in any apparent way. It seems it jumped from [4] to [6] directly.

Table 7: Example of a Misinterpretation and Wrong Conclusion due to False Premise. The complexities of this scenario are interpreting Marla's intent to remain indefinitely in Colorado by assessing her open-ended plans and lack of definite departure; distinguishing between temporary residence and establishing a new domicile by considering factors like her enrollment in a two-year program and a six-month lease against her willingness to stay in Colorado long-term; and applying domicile principles amid uncertainty.

Legal Context:
The basic standard for removal jurisdiction is simple and sensible: a defendant sued in state court may remove a case to federal court if it could have been filed originally in federal court. The rationale for removal is to give both parties access to federal court for cases within the federal court's jurisdiction. So it makes sense that the standard for removing a case to federal court should mirror the standard for filing one in federal court to begin with. So, 28 U.S.C. §1441(a) provides that "any civil action brought in a State court of which the district courts of the United States have original jurisdiction" may be removed to federal court. Consequently, the first question to ask in analyzing a removal problem is whether the claim asserted by the plaintiff fits into some category of federal jurisdiction. Is it a diversity case, or an arising-under case? A case between a citizen and foreign citizen, or some other case provided for in Article III, §2? If not, it can't be removed. If it is, then, with a few minor exceptions, it can be. Here's a first example to illustrate this relation between original and removal jurisdiction.

Question:
Castor Chemical Company, a California company, sues Pollux, from California, in state court for breach of contract, claiming that Pollux agreed to manufacture a drug for Castor and then reneged. Pollux answers the complaint, admitting that he refused to deliver the drug, and offering as an affirmative defense that he learned, after making the contract, that the drug was covered by a patent, so that, under federal patent law, he could not manufacture it without a license to do so. After answering, Pollux removed the action to federal court.

Options:
A. The action is properly removed, because at the time he was removed, Pollux had raised a federal law issue in his answer.
B. The action is properly removed, because the case "arises under" federal law within the meaning of that phrase in Article III, §2.
C. The action is not properly removed, because it is a federal question case, not a diversity case.
D. The action is not properly removed, because the case does not "arise under" federal law within the meaning of 28 U.S.C. §1331.

Human Annotator Analysis of Reasoning Chain 2:
1. The first step in determining whether a case can be removed to federal court is to identify whether the claim fits into any category of federal jurisdiction. In this case, the claim is for breach of contract between two California companies. (Factual Hallucination) - The breach of contract is between a person and a company.
2. The federal question jurisdiction, as per 28 U.S.C.§1331, applies to civil actions arising under federal law. In this case, Pollux has raised a federal law issue in his answer by asserting that he could not manufacture the drug due to a patent. (Correct)
3. However, the removal jurisdiction, as per 28 U.S.C. §1441(a), applies to civil actions brought in state court that have original jurisdiction in federal courts. (Correct)
4. In this case, the claim is not originally brought in federal court, but rather in state court. Therefore, even though a federal law issue has been raised, the case cannot be removed based on federal question jurisdiction alone.
(Misinterpretation) - The actual reason for why the case can not be removed to the federal court is that the plaintiff has invoked a state law tort in the claim and hence a federal court has no jurisdiction on this case even though the defendant raises a federal law issue. The correct standard is whether the plaintiff's original claim arises under federal law, not whether a federal issue is raised in the defense.
5. The correct answer is D: The action is not properly removed, because the case does not arise under federal law within the meaning of 28 U.S.C.§1331. (Correct Conclusion from False Premises)

(Both entities in same state jurisdiction ->Here, defendant is raising the federal jurisdiction claim-> Motley's 'arise under' won't apply (it only applies when plaintiff does that) ->Case can not be removed to Federal court)

Table 8: Example of Factual Hallucination and Misinterpretation. The complexities of this scenario would include: applying the well-pleaded complaint rule to determine whether a case arises under federal law; distinguishing between state law and federal law claims for removal purposes; and recognizing that federal defenses do not confer federal jurisdiction, as established in the Mottley rule.

Legal Context:
Let's start with the role of the United States Constitution in defining the limits of a court's power to subject a defendant to jurisdiction. The Fourteenth Amendment bars a state from depriving a person of life, liberty or property "without due process of law," that is, without a basically fair procedure. If it's a court that's doing the depriving—by entering a judgment against a person and forcing her to pay it—basic fairness requires that the defendant have some relationship to the state where the court sits that will make it fair to conduct the litigation there. In civil procedure terms, that means that the court must have "a basis to exercise personal jurisdiction" over the defendant. In the major personal jurisdiction cases, such as International Shoe, World-Wide Volkswagen, Daimler, Asahi, Burnham v. Superior Court, and Bristol-Myers Squibb, the United States Supreme Court has provided some guidance as to the types of relations to a state that will support the exercise of jurisdiction. Some relations that satisfy due process under the Fourteenth Amendment include domicile in a state,1 being "at home" in a state,2 minimum contacts that give rise to the claim,3 and service of process on an individual in the forum state.4 That's not an exhaustive list, and of course there are refinements and ambiguities, but the point is that the Supreme Court has upheld certain relations to a state as sufficient under the Fourteenth Amendment to support personal jurisdiction over a defendant there. However, it is not enough that a defendant has a contact with the state that is sufficient under the Due Process Clause to allow the state to exert jurisdiction over him. The state must also have a statute that authorizes the court to exercise that jurisdiction. This is confusing; if the exercise of jurisdiction is constitutionally permissible, why does the legislature have to authorize it in a statute too? The reason is that the Fourteenth Amendment does not actually confer jurisdiction on courts. It sets limits on the exercise of jurisdiction. ("No state shall deprive any person of life, liberty, or property without due process of law.") It says to the states, "go ahead, define your courts' jurisdiction as you like, but in doing so, don't overreach." **...Truncated**

Question:
Perrone, an Acadia resident, sues Margules, a Montanan, in an Acadia state court, for defamation. He claims that Margules called Acme Corporation, an Acadia company that was planning to hire Perrone, and made false and defamatory statements about his professional abilities that led Acme to decide not to hire him. Margules placed the call from her office in Montana. Assume that Acadia has a long-arm statute that authorizes personal jurisdiction over a defendant who "causes personal injury in this state by a tortious act or omission while physically present within this state." Assume that this is the only statute that might authorize the Acadia court to exercise personal jurisdiction over Perrone for this claim.

Options:
A. The Acadia statute authorizes the court to exercise jurisdiction over Margules in this case. However, it would be unconstitutional under the Fourteenth Amendment for the court to exercise jurisdiction over her on the facts of this case.
B. The Acadia statute does not authorize the court to exercise jurisdiction over Margules in this case. In addition, it would be unconstitutional under the Fourteenth Amendment for the court to exercise jurisdiction over her on the facts of this case.
C. The Acadia statute does not authorize the court to exercise jurisdiction over Margules in this case. However, it would be constitutional under the Fourteenth Amendment for the court to exercise jurisdiction over her on the facts of this case.
D. Although the Acadia statute does not authorize the court to exercise jurisdiction over Margules in this case, the court may do so because the claim arises out of her deliberate contacts with Acadia that gave rise to Perrone's claim.

Human Annotator Analysis of LLM generated Reasoning Chain:
[1] The first step is to determine whether the Acadia statute authorizes the court to exercise jurisdiction over Margules in this case. The statute in question authorizes personal jurisdiction over a defendant who causes personal injury in Acadia by a tortious act or omission while physically present within the state. (Correct)

[2] In this case, Margules is not physically present in Acadia when she allegedly committed the tortious act. She made the defamatory statements from her office in Montana. (Correct)

[3] Therefore, based on the text of the statute, it does not appear to authorize the court to exercise jurisdiction over Margules in this case.(Correct)

[Final Answer: B. The Acadia statute does not authorize the court to exercise jurisdiction over Margules in this case.](Wrong Conclusion from Incomplete Premises) - Making a decision based on an incomplete set of information, without considering all necessary legal aspects before jumping to the conclusion. Here the validity of the due process clause is not carefully examined and a premature conclusion is drawn.

However, it's important to note that even if the statute does not explicitly authorize jurisdiction in this case, a court may still exercise jurisdiction if the defendant has sufficient contacts with the forum state such that the exercise of jurisdiction comports with due process. This is known as the "minimum contacts" standard, which is a constitutional requirement separate from the statutory authorization. In this case, Margules' contact with Acadia through her defamatory statements may be sufficient to establish minimum contacts, but that analysis is beyond the scope of this question. (Misinterpretation) - No, that analysis is not outside the scope of this question!, in fact the expert has used "minimum" contact (instead of physical contact) to narrow the option to C. Margules' contact with Acadia through her defamatory statements is indeed be sufficient to establish minimum contacts and constitutionally

(Due Process Clause allows jurisdiction and is not violated but state statute is not satisfied)

Table 9: Example of Wrong Conclusion from Incomplete Premises. (The legal context has been truncated to save space.)

Legal Context:
The Due Process Clause of the Fourteenth Amendment prohibits a state from depriving a person of life, liberty, or property without due process of law. When a court enters a civil judgment against a person, it begins the process of taking the person's property (usually in the form of money). Consequently, the due process clause requires the court to use a fair procedure in entering judgment. Certainly, one component of a fair procedure is to tell the defendant that the court is going to adjudicate her rights. Hence, the Due Process Clause requires a court to use a constitutionally adequate means of notifying the defendant that a lawsuit has been commenced against her. Typically, it is the plaintiff who does the legwork of serving process on the defendant. The statutes or court rules in every state contain detailed provisions governing how this notice of a lawsuit is provided to the defendant. Court rules may authorize various means of serving process. The most obvious is to deliver the initiating papers in the case to the defendant in person, called "personal service of process." Other methods are often authorized as well, though they are less certain to actually inform the defendant about the suit. The service rules may provide, for example, that the papers may be left with someone at the defendant's home or place of business or slipped under the door. Some authorize service to the defendant by certified or ordinary mail. In some circumstances, court rules or statutes may authorize service by publication, by printing a notice of the suit in the legal notices in the local newspaper for several weeks. Such publication notice is sometimes referred to as "constructive service," as opposed to actual delivery of the papers to the defendant. Which of these methods is constitutionally adequate under the Due Process Clause? Justice Jackson's wise opinion in Mullane v. Central Hanover Bank & Trust Co., 339 U.S. 306 (1950) does not provide a mechanical answer to that question, but instead establishes broad standards for constitutionally adequate notice. Mullane involved a device called a common trust fund, which allowed a bank to pool a group of small trusts into one common fund, which could be managed more efficiently. Income (or losses) from the trust would then be shared by each smaller trust in proportion to their contributions to the fund. Periodically, the trustee of the fund would file accounts with the court. If the court approved the accounts, its judgment would bar investors from suing the bank for mismanagement. Since this judicial accounting could deprive trust beneficiaries of property—the right to sue the bank for mismanagement—due process required adequate notice to them of the proceeding. The problem in Mullane was that it was difficult to ascertain who all the beneficiaries were. Beneficiaries currently entitled to receive income from the trust were known, and the bank had their addresses in its files. But most trusts have contingent beneficiaries, such as the children or heirs of a beneficiary if the primary beneficiary dies. The bank might not know who these beneficiaries were, and they could change periodically. Perhaps, with a large investment of time by lawyers and investigators, the bank could identify most them, but doing so would eat up the profits of the common trust fund.**...Truncated**

Question:
In Mullane v. Central Hanover Bank & Trust Co., the Supreme Court held that

Human Annotator Analysis of LLM generated Reasoning Chain:
[1] The case of Mullane v. Central Hanover Bank & Trust Co. involved a common trust fund where the bank pooled multiple trusts into one fund for more efficient management. The trustee would file accounts with the court, and if approved, the judgment would bar investors from suing the bank for mismanagement, potentially depriving them of their property right to sue. The issue was who the bank had to notify about the suit and how. (Correct)
[2] The Supreme Court held that due process requires notice reasonably calculated to apprise interested parties of the pendency of the action and afford them an opportunity to present their objections. What is reasonable in one circumstance may not be in another. In Mullane, several factors reduced the risk of error if notice failed to reach all beneficiaries: the appointment of a guardian to represent their interests, and the likelihood that many beneficiaries would receive notice and share the interests of all beneficiaries in the settlement of the accounts. (Correct)
[3] Based on the information provided in the case, the Court did not require that each individual with an interest in the fund be given personal notice of the action. Instead, the Court focused on the reasonableness of the notice given under the circumstances.(Correct)
[4] Therefore, the correct answer is D: due to the large number of persons whose interests might be affected by the proceedings, notice by publication was sufficient on the facts of the case. (Wrong Conclusion from Incomplete Premises) - the LLM assumed on its own that notice by publication is sufficient, whereas the court ruling in Mullane, even though focusing on the reasonableness of the notice given in such circumstances, never mentioned that notice by publication was the right way! The ruling went for representation by Guardianship and Common Interests representation) Step [4] simplifies the Court's nuanced position into a broad and incorrect generalization. While the Court did acknowledge the use of publication notice as part of the process, it did not conclude that publication alone was sufficient under the circumstances of this specific case.

[Final Answer: D]

Table 10: Example of Wrong Conclusion from Incomplete Premises. (The legal context has been truncated to save space.)

Legal Context:
The relation between statutory diversity jurisdiction and the constitutional grant Article III, §2 of the Constitution creates diversity jurisdiction in very general terms. It does not expressly state whether there must be "complete diversity," or whether diversity might be satisfied if some plaintiffs are diverse from some defendants. Strawbridge v. Curtiss, of course, held that all defendants must be diverse from all plaintiffs, but for many years it was unclear whether Strawbridge interpreted Article III or interpreted the statute by which Congress conveys diversity jurisdiction to the lower federal courts. If Strawbridge interpreted the language "citizens of different states" in Article III, the complete diversity rule would be constitutionally required, and could not be changed **...Truncated**

Question:
A diversity case cannot be heard in federal court unless the amount in controversy is at least

Options:
A. $75,000, counting interest and costs as well as the plaintiff's damages.
B. $75,000, not counting interest and costs.
C. $75,000.01, counting interest and costs.
D. $75,000.01, exclusive of interest and costs.

Human Annotator Analysis of LLM generated Reasoning Chain:
[1] The legal context explains that the amount-in-controversy requirement is a condition imposed by Congress on the exercise of diversity jurisdiction in federal courts, which is not present in the constitutional grant of diversity jurisdiction in Article III, §2. (Correct)

[2] The question asks for the minimum amount in controversy that must be present for a diversity case to be heard in federal court. (Correct)

[3] Based on the legal context, the constitutional grant of diversity jurisdiction in Article III, §2, does not contain any monetary restriction. Therefore, there is no minimum amount in controversy required for a case to be considered a diversity case under the Constitution. (Irrelevant Premise) - The question is about the statutory requirement, not the constitutional one.

[4] However, the federal courts derive their jurisdiction from Congress, and Congress has imposed an amount-in-controversy requirement as a condition for the exercise of diversity jurisdiction. (Correct)

[5] Therefore, the correct answer is [B]: $75,000, not counting interest and costs. This is the minimum amount in controversy required for a diversity case to be heard in federal court under the statutory grant of jurisdiction (Wrong Conclusion from False Premises) - Mistral doesn't fully understand the terms of the legal context, which states "the amount in controversy exceeds the sum or value of $75,000, exclusive of interest or costs" not equals or exceeds, meaning the value must be greater than, not equal to, $75,000.

Table 11: Example of an Irrelevant Premise. (The legal context has been truncated to save space.)