FIMD Reasoning Evaluation Framework: Detecting the Deficiencies in Complex Legal Reasoning of Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) must demonstrate human-level reasoning capabilities to facilitate the broad adoption of machine learning in the legal industry. In pursuit of this goal, we introduce TortBench,[1] a dataset for legal reasoning that contains a collection of court judgments on tort cases annotated by a legal expert with summaries of legal explanations. We demonstrate how to formulate natural language explanation tasks to enhance the adoption of LLMs in the law domain. We test the reasoning capabilities of the most advanced LLMs and report the most frequent problems in their reasoning abilities. We introduce a novel framework for detecting limitations in LLM legal reasoning, flagging critical errors that may lead to harmful consequences along with novel metrics for benchmarking of reasoning capabilities. Our framework provides a foundation for future benchmarking and the continued improvement of legal reasoning in LLMs.

1 Introduction

In high-stake domains such as law, understanding the true rationales behind every prediction or recommendation is crucial. Without grasping the root causes of recommended actions, the implementation of automated systems in the legal industry remains doubtful. Legal cases, representing complex legal reasoning, are extensive documents. Consequently, legal firms often employ numerous paralegals and junior lawyers to synthesise and analyse these documents along with court practices. Detecting key points and effectively summarising case descriptions to make them relevant and understandable are critical aspects of legal work, where omissions can lead to severe consequences. To generate pertinent summaries of legal texts, one must possess knowledge of logical reasoning, an understanding of legal principles and laws, and a human-level understanding of the connections between them.

Large language models (LLMs) have demonstrated exceptional reasoning capabilities, competing for top performance metrics. For instance, the technical report for GPT-4 highlighted its high reasoning abilities, being in the 88th percentile on the LSAT exam (Achiam et al., 2023). GPT-4o was also capable of addressing the open-ended Multistate Essay Exam (MEE) and Multistate Performance Test (MPT) components, which demand robust reasoning capabilities (Katz et al., 2024). Meanwhile, Gemini has reported cutting-edge performance on tasks involving understanding and reasoning, positioning it alongside other advanced models in the field. Gemini Ultra achieved an accuracy of 90.04% on the Multi-task Language Understanding (MMLU) benchmark, surpassing human expert performance, which the benchmark authors gauged at 89.8% (Team et al., 2023).

Our objective is to evaluate the reasoning capabilities of advanced LLMs on legal texts specifically. To facilitate this, we first provide the TortBench dataset, a dataset of 143 complex legal texts annotated by human expert with concise summaries of legal reasoning, providing an overview of the key points and logic used in a legal decision. This dataset enables the community to assess and compare the results of these models against human analysis. Human-annotated summaries enhance the review process and allow the exploration of automated evaluation to assess the performance of LLMs.

We introduce a novel domain-specific reasoning framework designed to systematically evaluate the legal reasoning capabilities of LLMs and identify the most frequent deficiencies that may constrain their applicability in legal practice. Furthermore, we demonstrate that even the most advanced LLMs remain significantly below the threshold required to perform legal reasoning at a level necessary for

[1] We will make this dataset publicly available upon acceptance.

professional legal work, underscoring the need for further refinement and domain-specific adaptation.

2 Related Work

Existing explanation generation approaches fall into three categories: rule-based, retrieval-based, and generative methods. Rule-based methods use predefined templates but require expert annotation and lack generalization across legal tasks. Retrieval-based methods extract relevant text from a corpus, ensuring predictable output but limiting expressiveness, especially when access to diverse case law is restricted. Generative methods leverage LLMs for explanation generation, offering more flexibility in producing explanations (Zhang et al., 2014; Wang et al., 2018; Chen et al., 2018; DeYoung et al., 2019).

One of the main challenges in explanation generation is dataset formation. In order to train the models to generate explanations, the dataset with rationales should be sufficient to make a prediction (DeYoung et al., 2019).

Most datasets used in natural language explanations are domain-agnostic. For example, The CoS-E dataset with natural language explanations for commonsense reasoning is built on top of CQA, a question-answering challenge targeting commonsense knowledge (Talmor et al., 2019). BoolQ is a dataset comprising of explanatory passages selected from Wikipedia with yes/no questions regarding these passages (Clark et al., 2019). There is a dataset that can be used for biomedical data (Lehman et al., 2019). Due to complexity of legal language, existing datasets are not suitable for generating legal explanations.

Research on legal language understanding has received more attention recently, although it remains highly dependent on publicly available datasets. In research on legal language understanding, significant attention was given to judgement prediction (Chalkidis et al., 2019; Malik et al., 2021; Medvedeva et al., 2021; Alghazzawi et al., 2022; Cui et al., 2022; Ma et al., 2021) and legal question answering (Do et al., 2017; Fawei et al., 2019; Kim et al., 2015; Kien et al., 2020; Zhong et al., 2020). Only few research papers are related to other legal tasks, e.g., legal text generation (Wu et al., 2020) or textual entailment.

Most existing legal datasets for natural language understanding are designed to solve standard language tasks such as classification, prediction, and question answering (Chalkidis et al., 2022) or specific practical legal tasks (Guha et al., 2024; Fei et al., 2023) .

Unlike general-purpose LLM benchmarks or existing legal task benchmarks, our methodology identifies unique reasoning deficiencies specific to legal applications and introduces novel evaluation techniques along with supporting dataset to demonstrate existing limitations and drive advancements in the field.

3 The TortBench Dataset

We introduce a dataset annotated with summaries of explanations written by a legal expert, specifically focusing on a sample of appellate court judgments from Caselaw, a dataset of U.S. courts concerning tort cases (President and of Harvard University., 2024). We call our explanation-augmented dataset TortBench, benchmark dataset for tort case reasoning. It contains 143 legal cases, each carefully annotated to provide a clear, accessible understanding of complex legal decisions. The structured annotations in our dataset are designed to elucidate the pivotal issues and facts of each case. The 'issue' annotation briefly describes in several sentences the main grounds of the case, presenting the core legal questions or disputes being adjudicated. This component is crucial, as it frames the legal context and focal points of the judicial decision-making process. Accompanying this, the Summary annotation provides a brief description of the material facts and the application of law to these facts. This part of the annotation links directly to the issues and presents a description that logically connects the factual circumstances of the case with the applied legal principles. For an example of the case text and the corresponding annotation, refer to Appendix 9.

The application of tort law involves analyzing complex fact patterns that often contain ambiguous elements that require significant interpretation and judgment. Each case typically presents a unique set of facts in which specific details can greatly influence the application of the law, making tort cases particularly suited for evaluating legal reasoning. Lawyers and judges in this field must exercise a high degree of reasoning when applying abstract legal principles to specific circumstances. The nature of tort law, with its intricate and variable fact patterns, makes it an ideal domain for developing and testing LLMs to understand legal cases and provide legal explanations.

4 Prompt Selection

Prompt selection is a critical step in experiments involving LLMs. This aspect is particularly pertinent to reasoning tasks, where the formulation of the question may influence the outcomes. We developed five variations of prompts that convey the same task, albeit phrased differently. To ensure consistency, the phrasing used for human annotation was also included as one of the prompts for comparative analysis (Question 5).

The primary objective of testing various prompts is to identify the most effective formulation for complex legal reasoning tasks. The goal is to generate an optimal output that encompasses a comprehensive summary of the case, coupled with a detailed description of the court's judgment, directly derived from the description of case issues. Each prompt includes references to the legal case, the main question, and a format guide. While the reference to the legal case and the formatting instructions remained the same across all prompts, the question is different for each prompt.

Our prompts have the following structure: Based on {text}, {question}, {answer format}.

Question formulations:

Question 1. Explain why the court came to this conclusion;

Question 2. Describe what happened in the court case, and what the final verdict was;

Question 3. Provide a clear summary of the main points and outcome of the court case;

Question 4. Give a concise and easily understandable overview of the court ruling, including the essential details;

Question 5. Generate a plain English summary —- account in plain English of what happened in the case.

These prompts aim to evaluate the models' ability to synthesize complex information into coherent and detailed summaries and analyses, reflecting the nuances and outcomes of legal cases.

Additionally, the initial experiments aimed to assess whether the structure of the prompt significantly affects the outputs on complex legal reasoning tasks. We applied the selected prompts to generate explanations using GPT-4o on five randomly sampled legal cases. We detected four frequently occurring problems in the model output that can be summarized as follows:

1. Misleading Generalization (MG) occurs when the model's output, although generally accurate, omits critical facts or legally significant details that are essential for forming a robust legal conclusion. Such omissions can skew the reasoning process and lead to conclusions that might not hold if all relevant information was considered. For example, AI Explanation (Figure 2) describes the facts as, "Montanez was told all lines were dead by his supervisor". This description, while correctly reflecting Montanez's statement, neglects to mention that this statement was contested by the supervisor, which the court recognized as a genuine issue of fact. This omission is significant, as it overlooks a pivotal aspect of the case, misleading the reader about the disputed nature of the facts.

2. Incomplete Disclosure (ID) is identified when the output from the model fails to comprehensively address all relevant legal issues or facts or law necessary for a thorough legal analysis, resulting in a partial analysis.

3. False Representation (FR) occurs when the model incorrectly portrays the facts or legal context of a case, leading to a misinterpretation of the legal scenario. In Figure 2, the AI explanation incorrectly portrays the court's decision as a final judgement on the duty of care and adherence to safety standards. In reality, the court determined that these were matters containing substantial conflicting evidence, necessitating a jury's evaluation. The court reversed the summary judgment, emphasising that these complex issues, particularly whether the defendant's actions constituted negligence, and if they breached safety codes, should be thoroughly examined and decided in a trial setting. This distinction is crucial, as it highlights the preliminary nature of the decision and confirms the jury's essential role in resolving these factual disputes, rather than a conclusive judgement.

4. Deficient Reasoning (DR) indicates a lack of logical or legal coherence in synthesising the facts and applying legal principles and rules. For example, in Figure 3, the AI-generated answer states that the Arkansas Supreme Court relied on testimony and the circumstances of the accident to conclude that the tractor driver's actions were the proximate cause of the accident. However, the connection between these elements and the final legal judgement appears insufficiently explained. Additionally, it does not elaborate on how the court addressed the defendants' plea of contributory negligence, which was mentioned in the output of the model as one of the issues. By not illustrating the reasoning

Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore (F1)
GPT-4o	0.03 ± 0.02	0.27 ± 0.05	0.09 ± 0.03	0.23 ± 0.05	0.84 ± 0.01
Gemini	0.02 ± 0.03	0.26 ± 0.06	0.08 ± 0.03	0.22 ± 0.05	0.84 ± 0.02
Llama-3.3	0.04 ± 0.03	0.30 ± 0.06	0.10 ± 0.04	0.24 ± 0.05	0.84 ± 0.01

Table 1: BLEU, ROUGE, and BERTScore results for GPT-4o, Gemini, and Llama-3.3.

behind the decision, especially in terms of how the evidence directly supports or contradicts the issues (such as contributory negligence and sufficient lookout), the summary fails to provide a logical bridge between the facts and the legal conclusions.

To evaluate the performance of various prompts, we conducted a human evaluation. The evaluator was tasked with identifying the frequency of the problems described above in the outputs. During the review process, the prompts were anonymised to ensure that the expert is not biased by the prompts. The results are in Table 3. Our research revealed that Question 5 consistently produced outputs with misleading generalisations across all test cases. For Question 4, the majority of outputs exhibited deficient reasoning. More than half of the responses to Question 3 contained an incomplete description of the issue. No consistent pattern regarding these problems was observed in the responses to Questions 1 and 2. Notably, Question 1 demonstrated a less frequent occurrence of problems compared to others, and thereby it was selected as the prompt for further experiments.

	MG	ID	ER	DR
Question 1	0	1	0	1
Question 2	1	1	0	1
Question 3	2	3	1	2
Question 4	2	1	3	4
Question 5	5	2	3	3

Table 2: Frequency of problems across different prompts

5 Experiments

Our primary objective is to evaluate and compare the legal reasoning capabilities of state-of-the-art LLMs using zero-shot learning. We investigate whether these models are capable of generating human-like explanations of complex legal cases with all essential details from simple prompts without any prior specific training on similar tasks. We test a models' ability to generate explanations using a prompt with Question 1, across the following configurations:

- **Gemini-1.5-pro-001**: Configured with default parameters: temperature = 0.9, top_p = 1.0, top_k = 32, candidate_count = 1, max_output_tokens = 8192.
- **GPT-4o**: Evaluated using its default settings.
- **Llama-3.3-70B-Instruct**: Evaluated using its default settings without restricting the number of output tokens.

6 Evaluation using Score Metrics

We evaluated the models using BLEU, ROUGE, and BERTScore metrics across a dataset of 143 examples. In Table 1, we present the mean scores along with their standard deviations for BLEU, ROUGE, and BERTScore for GPT-4o, Gemini and Llama-3.3.

Our findings reveal a significant discrepancy between the BLEU and ROUGE score and BERTScore, suggesting that BLEU and ROUGE are not adequate metrics for this task. The mean BERTScore for Gemini, GPT-4o and Llama-3.3 demonstrated the same value, with a small standard deviation, indicating that these models are comparable in generating explanations for complex legal reasoning questions (Table 1).

For a more granular and practical approach to evaluating LLMs, we introduce a novel framework designed to systematically assess their performance in legal reasoning tasks.

7 FIMD Evaluation Framework

Human evaluation is challenging in the legal domain due to the limited availability of qualified human evaluators and the fact that the interpretation of laws is often shaped by the goals of the party advocating a particular position. To address these challenges, we propose an automated framework for evaluating text generated by large language models (LLMs) – FIMD Evaluation Framework.

This automated approach mitigates the influence of subjective biases and ensures a consistent and objective evaluation of generated legal content. Furthermore, it provides a solution to the scarcity of

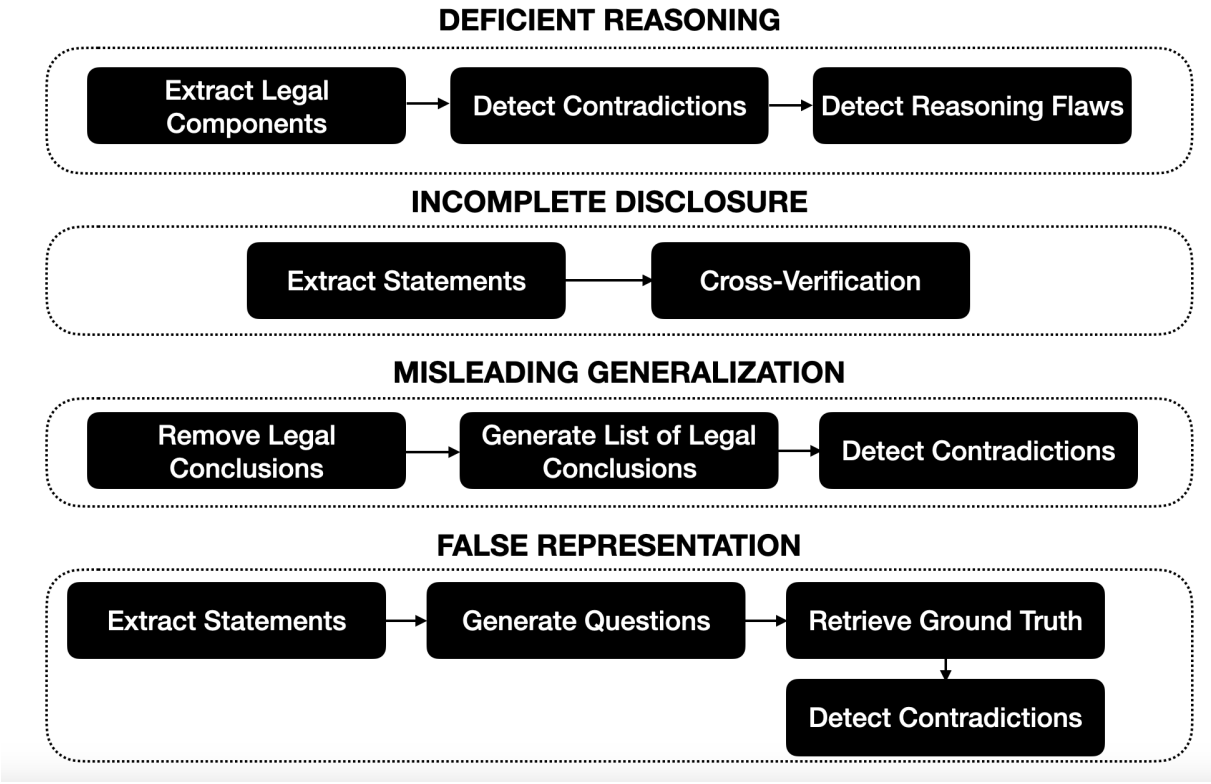

Figure 1: FIMD Reasoning Evaluation Framework.

qualified human evaluators, who are often required to possess extensive expertise in law and related domains. By automating the evaluation process, our framework can significantly reduce the time and resources needed for thorough reviews, while maintaining the reliability and accuracy necessary for legal analysis.

The proposed methodology is designed to systematically detect instances of Misleading Generalisation (MG), Incomplete Disclosure (ID), False Representation (FR) and Deficient Reasoning (DR) in generated legal texts by employing a structured combination of contextual analysis, logical inference, and advanced prompt engineering methods.

7.1 Identifying Misleading Generalization (MG)

The proposed methodology is divided into three primary stages: (1) removal of legal conclusions from the generated output, (2) generation of a list of possible legal conclusions from the generated issues and summary, and (3) detection of contradictions between the original conclusions and the generated conclusions.

Removing Legal Conclusions. The first stage aims to isolate the factual and legal principles embedded in the generated output, ensuring that any pre-existing legal conclusions are effectively removed. To achieve a high level of granularity, especially in cases where conclusions are not easily identifiable through keywords or linguistic markers and may be obscured by varying phrasing, a specially designed prompt is employed. This prompt instructs the model to split the legal text into two distinct components: (1) core information containing the facts and legal principles, and (2) the final court conclusion.

The extracted core information is stored separately from the identified conclusion, ensuring a clear delimitation between descriptive content and inferential statements. This stage is critical to maintaining the neutrality and objectivity of the analysis, as it prevents subsequent steps from being influenced by prior interpretations or biases.

Generation of Possible Legal Conclusions. The second stage aims to derive a comprehensive set of possible legal conclusions from the neutralized content. To accomplish this, a Language Model (LLM) is employed to infer potential conclusions based on the provided summary of facts and laws.

A structured prompt is crafted to instruct the LLM to analyze the input text and enumerate all plausible legal conclusions. The prompt empha-

sizes the need to base these conclusions strictly on the factual and legal descriptions provided, avoiding any unsupported or speculative inferences. The LLM processes the prompt and returns a list of inferred conclusions. This step ensures that a broad spectrum of potential interpretations is considered, capturing nuances that may lead to misleading conclusions due to the omission of legally significant information in the generated summary.

Contradiction Detection. The final stage involves comparing the generated conclusions against the original conclusions to identify any contradictions. Contradictions are defined as direct or implicit inconsistencies between the conclusions derived from the LLM and those articulated in the generated summary.

A second prompt is designed to instruct the LLM to perform a comparative analysis. This prompt includes the original summary and the generated conclusions, with explicit instructions to identify and explain any contradictions.

Misleading Generalization Score The Misleading Generalization Score measures how often a model generates misleading conclusions. It is calculated as follows:

$$\text{MG Score} = \frac{\text{Cases with Misleading Conclusions}}{\text{Total Number of Cases}} \quad (1)$$

A higher score indicates a greater tendency for the model to produce misleading legal interpretations.

7.2 Identifying False Representation (FR)

The proposed methodology is divided into four primary stages: (1) Statement Extraction, where factual and legal statements are systematically extracted from the generated output; (2) Question Generation, which involves creating a list of verification questions based on the extracted statements; (3) Ground Truth Retrieval, where answers to the questions from the previous step are extracted from original court judgement; and (4) Contradiction detection, which detects discrepancies between the generated output and the ground truth.

Statement Extraction The first stage focuses on Statement Extraction, where all factual and legal assertions are systematically identified and extracted from the generated output. This process isolates every claim made by the model, whether it pertains to facts or legal principles, ensuring that no component of the output is overlooked. By compiling a detailed list of these statements, this stage sets the foundation for subsequent verification and comparison.

Question Generation In the second stage, the question generation process is employed to convert the extracted statements into verification questions. For each factual or legal assertion from the previous step, a corresponding question is designed to assess its accuracy. For instance, if the extracted statement asserts, "The contract was signed under duress," the generated question might ask, "Was the contract signed under duress? What are the grounds?" This systematic approach ensures a thorough and precise evaluation of all statements extracted from the generated output.

Ground Truth Retrieval The third stage, Ground Truth Retrieval, involves sourcing answers for each question formulated in the previous step. These answers, referred to as Ground Truth, are derived from the original court judgement. This stage is critical to establish an objective benchmark against which the extracted statements will be evaluated, ensuring the integrity and reliability of the assessment process.

Contradiction Detection The final stage compares the statements from the generated output with the corresponding ground truth answers to identify any discrepancies. Each identified discrepancy is analyzed and categorized based on the level of agreement or divergence. Discrepancies are classified as either full discrepancies, where the generated output completely diverges from the ground truth, or full alignments, where the generated output and the ground truth are in complete agreement. This classification provides a clear framework for evaluating the accuracy and reliability of the model's output.

False Representation Score The False Representation Score evaluates how frequently a model generates outputs that contradict established legal facts:

$$\text{FR Score} = \frac{\text{Contradictory Statements}}{\text{Total Statements}} \quad (2)$$

A higher score indicates a greater risk of the model misrepresenting legal facts, which can be particularly problematic in legal applications.

7.3 Identifying Incomplete Disclosure

The proposed methodology is divided into two stages: (1) Statement Extraction, where all statements of law and facts are systematically extracted from both the human-provided annotation and the generated output to compile these into structured lists; and (2) Cross-Verification, which involves cross-referencing the extracted statements from the human annotation with those in the generated output, using prompt engineering to identify omissions, discrepancies, and misalignments in the generated output.

Statement Extraction This stage focuses on extracting all relevant statements of law and facts from the human-provided annotation and the generated output. To achieve this, specially crafted prompts are employed. The extracted statements are then compiled into structured lists that serve as the foundation for further analysis. This systematic approach ensures that both the human-provided annotation and the generated output are parsed thoroughly and consistently, enabling direct comparison.

Cross-Verification In this stage, the extracted lists of statements from the human-provided annotation are compared with those from the generated output to detect omissions or misalignments. Prompts guide the model to identify omissions. This process ensures that all critical legal and factual aspects are accounted for, providing a systematic framework to evaluate and enhance the comprehensiveness of the generated output.

Incomplete Disclosure Score The Incomplete Disclosure Score quantifies how much essential legal information is missing in the generated output:

$$\text{ID Score} = \frac{\text{Omitted Statements}}{\text{Total Statements in Annotation}} \quad (3)$$

A higher score suggests that the model fails to include key legal facts, resulting in incomplete explanations.

7.4 Identifying Deficient Reasoning

The proposed methodology is divided into three stages: (1) Logical Component Extraction, where the facts, cited laws, and reasoning chains are systematically extracted from the generated output and organized into structured elements; (2) Ground Truth Comparison, which involves cross-referencing these extracted components with a human-provided reference framework to identify inconsistencies; (3) Evaluation of Reasoning Flaws, where logical gaps, misapplications of legal principles, and unsupported conclusions are detected and categorized.

Legal Component Extraction In the first stage, the reasoning process of the generated output is deconstructed into its essential legal components. This involves systematically breaking down the output into three key elements: facts, laws cited, and application of laws to facts. Facts refer to the relevant factual information that the generated output identifies and presents as part of its reasoning. This step ensures that all significant contextual details are captured accurately. Next, the laws cited are extracted, which involves documenting the legal principles, statutes, or regulations that are referenced in the output. Finally, the application of laws to facts is mapped out, outlining how the model uses the cited legal principles to form logical connections with the identified facts. In this stage, prompt engineering is employed in this stage to structure queries and guide the extraction process, ensuring that all relevant components are identified systematically and accurately.

Ground Truth Comparison Following the extraction of logical components, the next stage involves a systematic comparison of these elements with a ground truth reasoning framework established by human experts. Specifically designed prompt plays a pivotal role in cross-referencing these components and highlighting the areas of divergence. Two key objectives guide this comparison: assessing the accuracy in law application and evaluating logical coherence. The accuracy assessment determines whether the generated output correctly applies the extracted legal principles to the relevant facts, while logical coherence focuses on whether the generated output synthesizes the laws and facts in a consistent with human annotation manner. This stage highlights discrepancies between the generated reasoning and the human benchmark, providing a robust foundation for further evaluation.

Evaluation of Reasoning Flaws The third stage focuses on identifying and categorizing deficiencies in the reasoning process. Common flaws include logical gaps, misapplication of laws, and unsupported conclusions. Logical gaps refer to instances where there are missing or unclear connections between facts, laws, and conclusions, leading to incomplete reasoning. Misapplication of laws

arises when the generated output incorrectly interprets or applies legal principles to the given facts, undermining the validity of its conclusions. Unsupported conclusions occur when the generated output draws conclusions without sufficient evidentiary or logical backing. Prompt engineering enhances this stage by enabling the generation of targeted queries to test specific reasoning paths.

Deficient Reasoning Score

The Deficient Reasoning Score (DRS) is calculated as the average number of reasoning flaws per case, defined as:

$$\text{ID Score} = \frac{\text{Total Reasoning Issues}}{\text{Total Number of Cases}} \quad (4)$$

A higher DRS indicates a greater presence of flawed reasoning in the model's generated legal explanations.

8 Evaluation of Legal Reasoning Deficiencies

To assess the reasoning capabilities of Large Language Models (LLMs) in legal contexts, we analyzed the sample of generated summaries using FIMD framework that capture deficiencies in model-generated legal reasoning capabilities.

Table 3: Evaluation of Legal Reasoning Deficiencies Across LLMs

Model	MG	ID	FR	DR
GPT-4o	0.9	0.32	0.05	5.8
LLaMA-3.3	0.7	0.42	0.09	5.4
Gemini-1.5	0.8	0.55	0.04	5.7

Our evaluations of large language models (LLMs) reveal notable variations in their ability to generate contextually appropriate summaries and explanations. GPT-4o exhibits the highest Misleading Generalization Score (0.9), indicating a tendency to summarize facts while omitting legally significant details, which can lead to misinterpretations by users relying on such summaries. Similarly, Gemini (0.8) and LLaMA-3.3 (0.7) also demonstrate a propensity for generalizations that deviate from precise legal reasoning. These findings suggest that while LLMs are capable of producing structured narratives, they frequently fail to maintain fidelity to the intricate details that define legal precision, compromising their reliability for professional applications.

In terms of Incomplete Disclosure, Gemini (0.55) exhibits the highest omission rate, failing to include essential legal statements within its generated explanations. LLaMA-3.3 (0.42) and GPT-4o (0.32) also demonstrated substantial omissions. The False Representation Score further highlights risks. While the overall False Representation Score is low, even a single instance of false representation can render an LLM unreliable in a legal context and lead to legal liabilities for a person who relies on such statements without verification. LLaMA-3.3 (0.09) exhibited the highest incidence of contradictions, with 25 false statements out of 284 extracted statements, making it the least trustworthy among the evaluated models. GPT-4o (0.05) and Gemini (0.04) demonstrated fewer contradictions but still pose significant risks.

9 Conclusion

In this study, we evaluated the performance of three leading LLMs, focusing on their ability to generate explanations for legal reasoning. Importantly, we introduced TortBench, a novel dataset comprising complex legal texts annotated with explanations by a legal expert. This dataset is specifically designed for evaluating reasoning models, providing a valuable resource for future research. Furthermore, we found that prompt selection plays a crucial role in influencing model outcomes, revealing issues such as misleading generalizations, false representations, deficient reasoning, and incomplete disclosure. These findings emphasize the need for further refinement in model training methodologies and prompt engineering to enhance the accuracy and reliability of generated explanations in complex reasoning law tasks. We introduced a novel framework for evaluating reasoning deficiencies in LLMs using automated metrics and demonstrated that, while generated outputs may appear coherent, they often contain critical deficiencies that pose significant risks in legal applications. Our findings reveal that even the most advanced LLMs frequently engage in misleading generalization, omit legally significant details, and introduce factual contradictions, making them unreliable for autonomous legal reasoning.

Limitations

Although the goal of this research was to assess the complex reasoning abilities of LLMs, our experiments are subject to limitations. First, the per-

formance of these models is constrained by their inherent limitations, including the reliability and accessibility of the APIs provided to interact with these models.

Second, although we employed a legal expert trained in law to annotate TortBench, human annotations are inherently subjective. These annotations can be influenced by individual perspectives, varying levels of expertise, and contextual factors, which can introduce bias into the evaluation process. Consequently, these annotations should be considered as one of several possible references to the ground truth, recognizing that legal interpretations can be diverse and that other valid interpretations may exist.

Finally, our study is limited to the specific legal reasoning tasks and datasets used. The generalization of our findings to other domains or legal contexts remains an open question and requires further investigation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Daniyal Alghazzawi, Omaimah Bamasag, Aiiad Albeshri, Iqra Sana, Hayat Ullah, and Muhammad Zubair Asghar. 2022. Efficient prediction of court judgments using an lstm+ cnn neural network model with an optimal feature set. *Mathematics*, 10(5):683.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. *arXiv preprint arXiv:1906.02059*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 World Wide Web Conference*, pages 1583–1592.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Junyun Cui, Xiaoyu Shen, Feiping Nie, Zheng Wang, Jinglong Wang, and Yulong Chen. 2022. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *arXiv preprint arXiv:2204.04859*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*.

Phong-Khac Do, Huy-Tien Nguyen, Chien-Xuan Tran, Minh-Tien Nguyen, and Minh-Le Nguyen. 2017. Legal question answering using ranking svm and deep convolutional neural network. *arXiv preprint arXiv:1703.05320*.

Biralatei Fawei, Jeff Z Pan, Martin Kollingbaum, and Adam Z Wyner. 2019. A semi-automated ontology construction for legal question answering. *New Generation Computing*, 37(4):453–478.

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*.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254.

Phi Manh Kien, Ha-Thanh Nguyen, Ngo Xuan Bach, Vu Tran, Minh Le Nguyen, and Tu Minh Phuong. 2020. Answering legal questions by learning neural attentive text representation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 988–998.

Mi-Young Kim, Randy Goebel, and S Ken. 2015. Coliee-2015: evaluation of legal question answering. In *Ninth International Workshop on Juris-informatics (JURISIN 2015)*.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume*

1 (Long and Short Papers), pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

Luyao Ma, Yating Zhang, Tianyi Wang, Xiaozhong Liu, Wei Ye, Changlong Sun, and Shikun Zhang. 2021. Legal judgment prediction with multi-stage case representation learning in the real court setting. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 993–1002.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripa Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation. *arXiv preprint arXiv:2105.13562*.

Masha Medvedeva, Ahmet Üstün, Xiao Xu, Michel Vols, and Martijn Wieling. 2021. Automatic judgement forecasting for pending applications of the european court of human rights. In *ASAIL/LegalAIIA@ ICAIL*.

The President and Fellows of Harvard University. 2024. Caselaw access project. *https://case.law*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. 2018. Explainable recommendation via multi-task learning in opinionated text data. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 165–174.

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-biased court's view generation with causality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 763–780.

Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 83–92.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jec-qa: a legal-domain question answering dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9701–9708.

A Example of Expert Annotation

Issue: Whether the trial court should have granted summary judgment, and whether material issues of fact existed. **Summary**: The trial court granted summary judgment in favor of all of the defendants, including the child driver of the truck from which the deceased victim fell, and the coach supervising the baseball training session at which the incident occurred. The appellate court found that there were material issues of fact that should have been put to the jury; however, the coach was rightly granted summary judgment. Whether the driver of the truck should have known, if exercising reasonable care, that the passengers were not inside the truck but were hanging onto the back of it was a material issue of fact. Therefore, the grant of summary judgment in favor of the defendant child driver was erroneous. Given that the baseball players had never disobeyed the coach before and were not prone to reckless behavior, there was no reason that the coach should have foreseen that they would disobey his direction to ride in the truck, as opposed to on the back of it. As a result, the injury was not reasonably foreseeable, and the grant of summary judgment in favor of the defendant coach was proper.

Case: Marquez ex rel. Estate of Marquez v. Gomez, 116 N.M. 626, 866 P.2d 354 (1991)

Human Annotation	AI Explanation
Issue: whether defendants owe a duty of care to the plaintiff who was injured when dismantling electrical equipment Summary: the plaintiff was employed by an electrical contractor as an electrician's helper, and had limited experience with electricity. The plaintiff was directed to take down electricity lines and poles at the site of the defendant oil well owner, including a 'secondary system' installed by defendant Cass. The plaintiff was told by his employer that the lines were dead, but some were in fact carrying electricity, and the plaintiff touched a live wire and was injured. Utility company the plaintiff claimed that the defendant electricity utility company, which installed the pole, but not the 'secondary system', owed a duty to maintain safe operation of its electricity lines. However, the defendant utility company did not have knowledge of the installation of the secondary system by defendant Cass, and also did not own the secondary system, but merely supplied electricity to it. Thus, although the defendant utility company has a duty to inspect its electricity lines and discover defects, this does not extend to secondary systems installed by third parties without its knowledge, so the defendant utility owed no duty to the plaintiff. defendant Cass, who installed the secondary system despite being an independent contractor, defendant Cass owed a duty to anyone who might foreseeably be endangered by allegedly defective performance of its work. The plaintiff, as an employee of another electrical contractor, might foreseeably have been injured, so defendant Cass owed him a duty (and thus summary judgement was improper). defendant Wolfson, owner of the oil well the Court of Appeals had held that defendant Wolfson owed a duty to the plaintiff because he was performing 'inherently dangerous' work. The Supreme Court held, however, that although employers of contractors owe some duties to 'others' for work the contractors perform, 'others' does not include independent contractors' employees, since that would be to impose greater responsibility on the employer for contractors' employees than for the employer's own employees (who are mandatorily covered by insurance). There is no reason the employer should be responsible for the employees of an independent contractor, since those employees are the responsibility of the contractor. Since defendant Wolfson retained no control over the work performed by the plaintiff, it had no duty to him.	Issue: The case revolves around the plaintiff, Montanez, who was working as an electrician's helper for Gary Electric. He was injured while attempting to dismantle a secondary electric system at an oil well owned by Wolfson, which had been installed by Cass. Allegedly, Montanez was told all lines were dead by his supervisor; however, while climbing a pole, he came into contact with a live wire and fell to the ground. The legality questioned here lies in determining which party was negligent and liable for Montanez's injuries. Summary: After analyzing each defendant's involvement and responsibilities in the case, the Supreme Court of New Mexico determined the following: - New Mexico Electric Service Company (Utility) could only be held liable if it had furnished electricity with knowledge of a defect in the system installed by Cass. Since there was no evidence of such knowledge, and the utility was not responsible for inspecting the system, it had no duty to Montanez. Thus, the trial court's summary judgment for the Utility was affirmed. - Cass, the company that installed the secondary system, had a duty of care to anyone likely endangered by faulty construction. This includes Montanez as an employee of another independent electrical contractor. Whether Cass had breached its duty of reasonable care or whether its construction of the secondary system proximately caused Montanez's injuries was for the jury to determine. Thus, the trial court's summary judgment for Cass was reversed. - Lastly, Wolfson, the oil well owner, had no duty of care to Montanez as an employee of an independent contractor. Wolfson merely hired independent contractor Gary Electric to dismantle an electrical system; therefore, it wasn't liable for employees working under that independent contractor. The trial court's summary judgment for Wolfson was affirmed. Finally, the issue of Montanez's contributory negligence was to be decided by a jury. In other words, the case was partly sent back for trial (with respect to Cass' liability) and partly upheld in favour of the defendants (the Utility and Wolfson).

Figure 2: Example of Misleading Generalization and False Representation

Case: Cobb v. Atkins, 239 Ark. 151, 388 S.W.2d 8 (1965)

Human Annotation	AI Explanation
Issue: Whether the defendant tractor driver was negligent in failing to keep a lookout and driving negligently and causing the death or injury of the plaintiffs. Summary: The plaintiff was in a car that attempted to overtake a tractor driven by the defendant in the same direction; the tractor swerved as it was overtaken and clipped the plaintiff's car, which car was overturned and lay on its side, where another car driven by a separate plaintiff collided with it. The plaintiff's wife and two others were killed. The defendant said that the plaintiff had failed to adduce evidence that supported the allegation that the defendant had failed to keep a lookout, but the trial court (affirmed by the Supreme Court) held that the testimony given by the plaintiff—that the driver of the plaintiff's car had used his horn, and that the defendant was driving at a high rate of speed—was sufficient for the jury to be given the opportunity to give a verdict. Similarly, there was evidence that the tractor had been driven in the middle of the road, even though the defendant said there was no such evidence. Finally, the evidence given by the plaintiff was sufficient evidence of negligence such that the jury was rightly given the opportunity to give a verdict. Thus, the evidence fairly supported the verdict of negligence given by the jury.	Issue: The case ensued after a traffic accident involving a farm tractor and two automobiles which resulted in the death of three people and injuries to three others. The defendants were the owner of the tractor, Clarence Cobb, and the tractor driver and his employee, Archie Earls. The plaintiffs alleged that Earls, served as Cobb's servant, was driving the tractor in a negligent and dangerous manner that led to the collision. The defendants denied any wrongdoing and pled for contributory negligence. Summary: The Arkansas Supreme Court had to decide whether there was actionable negligence on the part of the tractor driver and whether the evidence presented was sufficient to support the allegations of the plaintiffs. The Supreme Court ruled in favor of the plaintiffs, affirming the trial court's decision and the verdicts against the defendants. It concluded that there was enough evidence presented to the jury, such as Atkins' testimony and the circumstances of the accident, that indicated faulty lookout and erratic driving by Earls which could have been the cause of the accident. The Court also ruled that the tractor crossing the center line to strike the Cox car could have been the proximate cause of the accident. Therefore, the court saw no merit in the defendants' appeals and upheld the verdicts and judgments against them.

Figure 3: Example of Deficient Reasoning

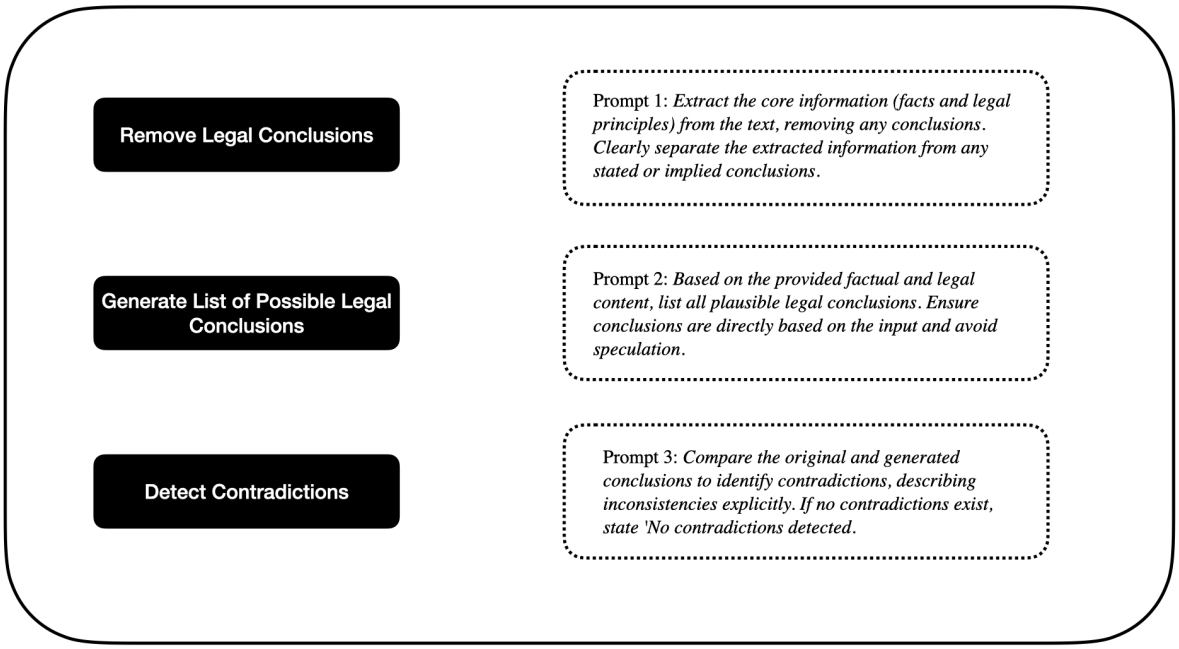

Figure 4: Identifying Misleading Generalization.

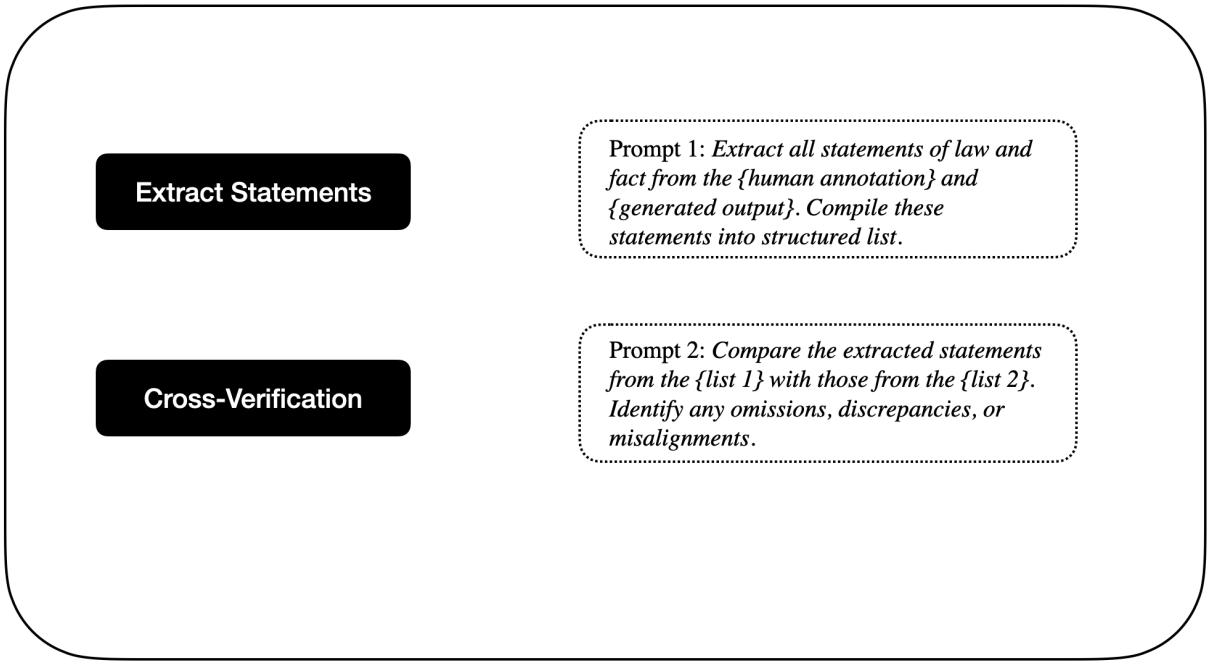

Figure 5: Identifying Incomplete Disclosure.

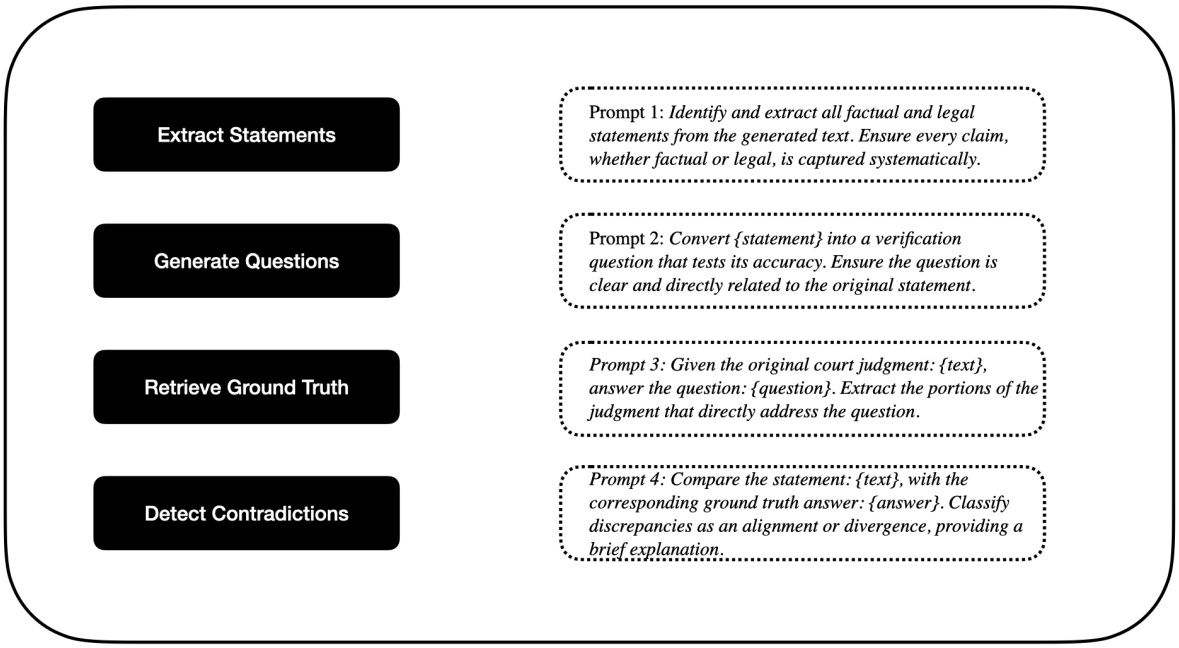

Figure 6: Identifying False Representation.

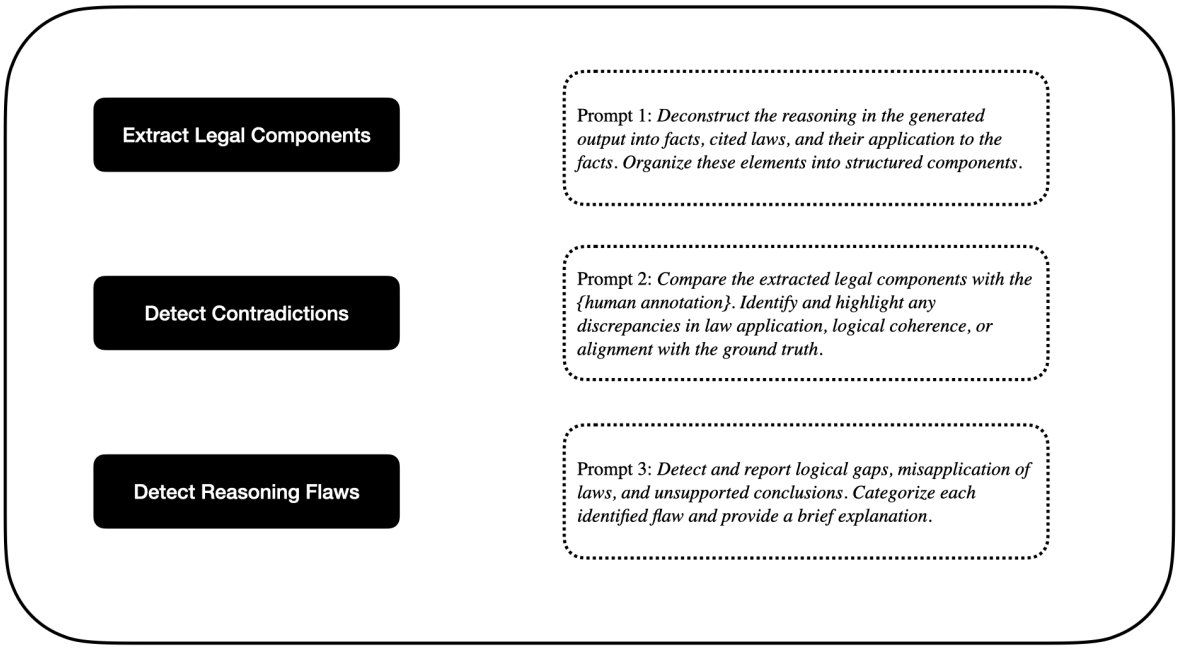

Figure 7: Identifying Deficient Reasoning.