# ENHANCING LLMS IN LEGAL JUDGMENT PREDICTION VIA NEURO-SYMBOLIC REASONING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Large Language Models (LLMs) often struggle with Legal Judgment Prediction (LJP) tasks, failing to maintain the rigorous logical consistency required for legal judicial decision-making despite their outstanding semantic capabilities. Existing methods relying on LLMs' reasoning capabilities remain prone to instability and hallucinations, and thus there is the absence of logically reliable and explainable methods for LLMs in LJP task. To fill this gap, we propose a novel neuro-symbolic approach that integrates an external logical solver to determine whether the conduct in the case fact constitutes a violation of specific law articles. Specifically, our approach utilizes an LLM to translate texts into symbolic representations, performs reasoning via an external solver to determine the logical consistency between the articles and facts, and interprets the output to ensure the final answer is both logically accurate and contextually readable. Experiments demonstrate that our method significantly outperforms both general and law domain-specific reasoning baselines.

## 1 INTRODUCTION

Legal Judgment Prediction (LJP) stands as an important task in the field of Artificial Intelligence and law, aiming to automatically predict judicial outcomes of whether the conduct in the case fact falls under certain articles in law (Zhong et al., 2020; Cui et al., 2023). Recently, Large Language Models (LLMs) have demonstrated remarkable performance on advancing various legal tasks such as legal text summarization, information retrieval, and element extraction (Chalkidis et al., 2020; Katz et al., 2024) due to their outstanding capabilities in natural language processing. Despite these impressive successes, the performance of LLMs remains limited in LJP tasks, which require complex and multi-step logical reasoning. While LLMs excel at capturing various semantic patterns, they often struggle to maintain the rigorous logical consistency required for judicial decision-making, leading to plausible-sounding but legally inaccurate predictions (Huang & Chang, 2023).

To enhance the performance of LLMs in the complex legal task LJP, previous general strategies that can be used include Chain-of-Thought (CoT) prompting (Wei et al., 2022), retrieval-augmented generation (RAG) (Lewis et al., 2020), and law domain-specific fine-tuning (Deng et al., 2024). However, these methods primarily rely on the inherent implicit reasoning capabilities of the LLM that often lack the logical consistency and interpretability, which is essential for the application in legal scenarios. As a result, the performance of LLMs in LJP task is often prone to instability and hallucinations, where the model generates factually conflicting statements (Ji et al., 2023). Currently, there is a notable absence of logically reliable and explainable methods for LLMs in LJP task, particularly for civil law systems like Chinese Criminal Law, which demands strict adherence to the logical reasoning rules and consistent mapping between the case facts and the legal articles.

To fill this gap, we propose a novel neuro-symbolic framework integrating a logical solver to determinate whether the conduct in the case fact constitutes a violation of certain articles in law. Specifically, we first utilize an LLM to translate all the relevant articles and case facts in natural language (NL) into symbolic language (SL). Based on these SL representations, we use LLMs to generate a query in corresponding SL for each *(article, case facts)* pair to determine *whether [the person's conduct in the case] satisfies [the charge in the law article]*. Then these symbolic formulations are input into a logical solver, which performs logical deduction to determine whether the charge in the article is logically consistent with the conduct in the case. Finally, the LLM interprets the solver's

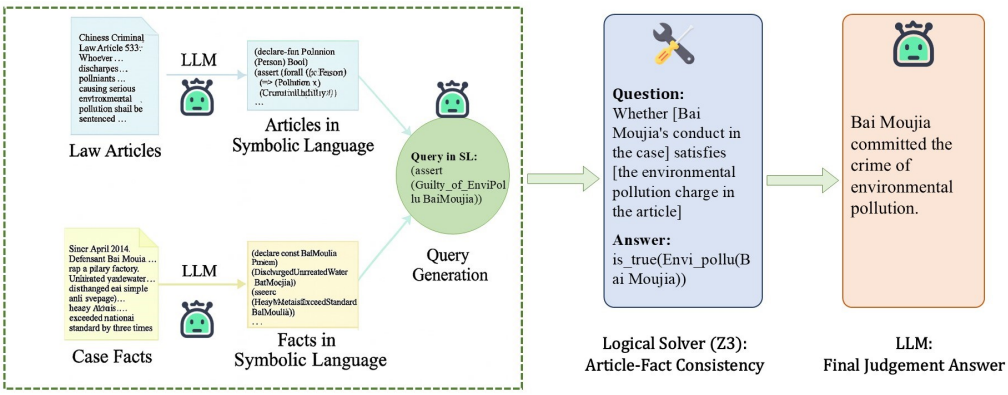


Figure 1: The overall framework of our neuro-symbolic method. The pipeline consists of four stages: (1) translating law articles and case facts from NL to SL via an LLM; (2) generating logical queries to verify if the conduct satisfies the charge; (3) determining the article-fact consistency via an external logical solver (Z3); and (4) interpreting the solver's output to generate the final judgment answer via the LLM.

SL output to generate the final judgment answer in NL, ensuring the LLM's output is both logically accurate and contextually readable. The contributions of this paper are summarized as follows:

- We introduce a novel perspective for improving LLMs performance in complex legal tasks such as LJP by harnessing their symbolic logical reasoning capabilities.
- We propose a neuro-symbolic approach for the LJP task that incorporates an external logical solver to reason and an LLM to interpret, ensuring both the logical consistency and textual readability of legal reasoning.
- We conduct experiments comparing our method with both general and law domain-specific baselines, demonstrating the superiority of introducing neural symbolic reasoning in the LJP task.

## 2 Related Work

### 2.1 Improving Performance of LLMs in Legal Tasks

Research on improving legal tasks performance of LLMs include three paradigms: domain-specific fine-tuning, RAG, and reasoning prompting (Cui et al., 2024). Fine-tuning injects professional knowledge via legal corpora to align models with correct legal knowledge and reasoning (Yue et al., 2023; Huang et al., 2023; Colombo et al., 2024). RAG mitigates hallucinations by retrieving external legal knowledge to ground generation (Louis et al., 2023; Fei et al., 2023). Prompt-based methods utilize CoT to mimic syllogisms (Jiang & Yang, 2023; Katz et al., 2024) or structured frameworks for interpretation (Savelka et al., 2023; Deng et al., 2023). Differing from previous works, we integrate an external logical solver, addressing the limitations of LLMs in maintaining logical consistency during the legal judgment requiring multi-step reasoning.

### 2.2 Enhancements in Logical Reasoning of LLMs

Logical reasoning improvements for LLMs generally follow solver-based, prompt-based, or fine-tuning approaches (Cheng et al., 2025). Solver-based methods translate NL into SL for symbolic reasoning via an external logical solver (Olausson et al., 2023; Ye et al., 2023; Ryu et al., 2025). Prompt-based strategies either elicit explicit NL reasoning chains (Wei et al., 2022; Yao et al., 2024; Zhang et al., 2024) or prompt LLMs to perform NL-to-SL translation, reasoning and verification processes (Liu et al., 2025; Xu et al., 2024; 2025). Fine-tuning internalizes LLMs logical reasoning abilities via synthetic datasets containing logical proofs or augmented corpora (Bao et al., 2024; Morishita et al., 2024; Feng et al., 2024; Wan et al., 2024; Jiao et al., 2024). Focusing on the LJP

task, we propose a neuro-symbolic framework integrating a logical solver to determinate the fact-article consistency.

## 3 LEGAL JUDGMENT PREDICTION TASKS

LJP is a fundamental legal task aiming to predict judicial results, such as applicable charges, based on the textual description of case facts (Zhong et al., 2020). This process requires the model to rigorously map the defendant's conduct to the constitutive elements of specific legal articles, rather than merely relying on semantic similarity. An example of LJP task from the CAIL2024 dataset) (Huang et al., 2024) is as follows, where the goal is to determine if the specific charge of "Theft" applies to the described behavior:

**Example from CAIL2024 Dataset**

**Case Fact:** On the evening of March 12, 2023, the defendant Zhang noticed that the victim Li had left his electric scooter unlocked outside a supermarket. Zhang took advantage of the situation to ride the scooter away. The scooter was later valued at 2,500 RMB.

**Relevant Article:** *Criminal Law of the PRC, Article 264 (Theft):* Whoever steals public or private property for a relatively large amount... shall be sentenced to fixed-term imprisonment...

**Prediction Task:** Does the defendant's conduct constitute the crime of Theft?
**Ground Truth:** True

## 4 PROPOSED METHOD

We introduce a neuro-symbolic framework for LJP task integrating a logical solver to determinate whether the conduct in the case fact constitutes a violation of certain articles in law. As shown in Figure 1, the method includes four stages: NL-to-SL translation for law articles and case facts, construction of logical queries, determining the consistency via an external solver, and final judgment interpretation via the LLM. This approach ensures the final output is both logically accurate and contextually readable.

### 4.1 TRANSLATING ARTICLES AND FACTS INTO SL

We first utilize an LLM to translate unstructured NL texts into SL representations compatible with Satisfiability (SAT) solvers (we employ Z3 here). Specifically, relevant legal articles and cases facts are parsed into SAT formulations involving implication and quantification. We instruct the LLM to maintain a consistent NL-to-SL mapping to ensure that identical concepts across articles and case facts are bound to the same symbolic representations.

### 4.2 QUERY GENERATION AND COMBINATION

After obtaining the symbolic representations, we construct a logical query to evaluate the consistency between the case facts and the law articles. Specifically, for each *(article, case facts)* pair, the LLM generates a targeted query in SL aimed at determining whether the defendant's conduct satisfies the constitutive elements of the charge (e.g., *is_true(Envi_pollu(Bai Moujia))*). This SL query combined with the previously generated SL representations of the article and facts, will be input into an external solver.

### 4.3 REASONING VIA THE LOGICAL SOLVER

The combined symbolic representations are input into an external logical solver (Z3) to perform rigorous deductive reasoning. The solver deterministically evaluates the satisfiability of the constructed query in SL based on the provided articles and facts in SL. It determines whether the case facts (as the premises) can logically entail the charge in the article (as the conclusion). The solver returns a precise boolean result (Satisfied/Unsatisfied), effectively validating the consistency between the article and the facts.

Table 1: Comparative evaluation of charge prediction accuracy (%) across five diverse datasets. We benchmark our proposed neuro-symbolic method against five prompting strategies (Zero-Shot, CoT, ToT, CoR, and LSL).

(a) Performance on DeepSeek-V3

| Dataset | Zero-Shot | CoT | ToT | CoR | LSL | Our Method |
|---|---|---|---|---|---|---|
| **TDSIF** | 69.61% | 68.63% | 70.59% | 71.00% | 75.00% | **95.00%** |
| **RPPID** | 54.08% | 57.14% | 52.04% | 60.00% | 56.00% | **89.47%** |
| **OALEC** | 67.00% | 61.00% | 63.00% | 83.00% | 77.00% | **100.00%** |
| **LawBench** | 38.00% | 37.00% | 35.00% | 51.00% | 41.00% | **78.79%** |
| **CAIL2018** | 76.00% | 79.00% | 77.00% | 82.00% | 80.00% | **100.00%** |

(b) Performance on GPT-4.1

| Dataset | Zero-Shot | CoT | ToT | CoR | LSL | Our Method |
|---|---|---|---|---|---|---|
| **TDSIF** | 66.67% | 74.51% | 78.43% | 79.00% | 65.00% | **85.00%** |
| **RPPID** | 54.08% | 57.14% | 60.20% | 72.00% | 54.00% | **82.11%** |
| **OALEC** | 65.00% | 62.00% | 67.00% | 75.00% | 72.00% | **98.00%** |
| **LawBench** | 36.00% | 34.00% | 38.00% | 45.00% | 34.00% | **73.74%** |
| **CAIL2018** | 70.00% | 68.00% | 83.00% | 92.00% | 74.00% | **100.00%** |

### 4.4 Final Answering via the LLM

In the final stage, the raw output from the logical solver is fed back into the LLM for interpretation. The LLM translates the solver's deterministic result into a NL judgment, providing a coherent and readable answer (e.g., "Bai Moujia committed the crime of environmental pollution"). This step bridges the gap between formal logic and human-readable reasoning, ensuring that the final output is not only logically accurate but also contextually appropriate for legal users.

## 5 Experiments

### 5.1 Experimental Setup

The experiments are conducted using two LLMs: **DeepSeek V3** and **GPT-4.1**. We evaluate the performance of our method on several LJP datasets against five widely used prompting baselines in legal tasks.

**Benchmark Datasets.**

- **TDSIF, RPPID, OALEC:** Three challenging subsets extracted directly from the CAIL2024 (Chinese AI and Law Challenge) (Huang et al., 2024) corpus.
- **LawBench (Subset) (Fei et al., 2023):** A subset curated from the LawBench corpus for comparison.
- **CAIL2018 (Subset) (Xiao et al., 2018):** A subset curated from CAIL2018 corpus, serving as a standard baseline.

**Evaluation Metric and Baseline Methods.** The primary metric used is **Accuracy (%)**—the percentage of cases where the LLM correctly predicts the final criminal charge (Ground Truth). We compare our method with the following five baseline reasoning strategies:

- **Zero-Shot:** Direct prediction without explicit reasoning.
- **CoT (Chain-of-Thought)** (Wei et al., 2022): Sequential reasoning leading to the conclusion.
- **ToT (Tree-of-Thought)** (Long, 2023): Exploring and evaluating multiple possible reasoning paths.
- **CoR (Chain of References)** (Kuppa et al., 2023): Reasoning supported by specific cited legal articles or principles.
- **LSL (Legal Search Logic)** (Martin et al., 2024): Simulating real-world legal research by retrieving information and drawing conclusions based on the search results.

### 5.2 Results Analysis

**Overall Performance and Superiority of Our Method.** Our method achieves the highest accuracy across all datasets and models, with performance gains ranging from 8% to over 40% compared to the best-performing baseline. Notably, the method achieves perfect accuracy (100.00%) for both DeepSeek V3 and GPT-4.1 on the OALEC and CAIL2018 subsets. This strong performance confirms that structuring the LLM's reasoning process through SL significantly enhances its ability to handle the complexity and nuances of criminal charge prediction.

**Analysis of Baseline Methods.** The reasoning-based baselines (CoT, ToT, CoR, LSL) generally outperform the Zero-Shot method, validating the need for explicit step-by-step processing in legal tasks. Among these, CoR (Chain of References) often shows a slight edge, suggesting that grounding the reasoning in specific legal provisions is a beneficial strategy. However, the plateau in accuracy observed across all baselines (with the maximum accuracy being $92.00\%$) indicates their susceptibility to flaws in free-form generation, particularly when facing tasks involving contradictory legal interpretations or complex fact patterns.

**Model Comparison.** The results show that DeepSeek V3 is generally competitive with, and in several instances outperforms (e.g., in Zero-Shot, CoR, and SAT on TDSIF and LawBench), GPT-4.1 when using the same reasoning technique. The most critical observation is that the performance boost conferred by our method is robust across different foundation models, consistently improving their performance in the LJP task.

# 6 Case Study

We present a detailed analysis of a complex legal case involving the concurrence of offenses. This case demonstrates how our method's reasoning via solver prevent the semantic distractions that frequently lead baseline reasoning methods (CoT, CoR, LSL) to incorrect legal characterizations.

## 6.1 Case Background and Legal Conflict

The case involves the unauthorized sale of counterfeit cigarettes. Under Chinese Criminal Law, this scenario presents a concurrence between the *Crime of Selling Counterfeit Registered Trademark Goods* (Article 214) and the *Crime of Illegal Business Operation* (Article 225).

**Case Facts:** From 2009 to 2012, Defendant Qiu purchased counterfeit "Ruan Zhonghua" cigarettes and sold $510$ cartons to Defendant Nan (Amount: $109,000$ RMB). Nan subsequently resold these to Wang Yi (Amount: $81,600$ RMB). Both defendants admitted to the facts. The cigarettes bore counterfeit registered trademarks. Crucially, tobacco is a state-controlled monopoly product.
**Ground Truth Charge:** Illegal Business Operation.

The legal ground truth prioritizes the violation of the state monopoly order (Article 225) over the trademark infringement (Article 214), when the prosecution threshold for Illegal Business Operation is met, which demonstrates the adherence to the principle of *lex specialis*.

## 6.2 Analysis of Baseline Failures

Baseline methods (CoT, LSL, CoR) consistently misclassified the case as the *Crime of Selling Counterfeit Registered Trademark Goods*. As shown in the box below, the Chain-of-Reference (CoR) method correctly identifies the monopoly status but commits a logical error by subjectively weighting the "prominence" of the trademark infringement over the monopoly violation. This illustrates a vulnerability to salient semantic features ("counterfeit", "trademark") at the expense of hierarchical legal principles.

**Baseline Error (Representative CoR Output):**
*Step 4 [Evaluation]:* The facts clearly confirm defendants sold "counterfeit Ruan Zhonghua cigarettes", directly infringing trademark rights.
*Step 6 [Reasoning Failure]:* **Although cigarettes belong to state-monopoly goods, the prominent illegality in this case lies in the infringement of trademark rights**, making the application of the criminal law provision protecting intellectual property rights more appropriate.
*Output Charge:* ⟨box⟩Crime of Selling Counterfeit Registered Trademark Goods⟨/box⟩

### 6.3 Analysis of Our Method Success

Our method correctly identifies the charge by enforcing a structured evaluation of object attributes. Unlike the free-form reasoning of baselines, the SAT solver is constrained to verify the `is_state_controlled` attribute. Once the object (cigarettes) is confirmed as a monopoly good and the act is unlicensed sale, the logic strictly maps to Article 225, effectively ignoring the distraction of the trademark issue which acts as a "distractor" feature in this context.

**SAT Method Success (Solver Trace):**
*1. Solver Selection:* Assert `has_action(Nan, sell, fake_cigarettes, 510, 160) == True`.
*2. Legal Attribute Check:* **Cigarettes are state-controlled monopoly goods (tobacco monopoly products)** as stipulated by national laws.
*3. Charge Determination:* Engaging in the purchase and sale of tobacco monopoly products without a license severely disrupts the national market monopoly management order. This act complies with Article 225.
*Output Charge:* ⟨box⟩Illegal Business Operation⟨/box⟩

## 7 Conclusion

In this paper, we proposed a novel neuro-symbolic framework to enhance LLMs' performance in LJP. By integrating an external logical solver to strictly determine the consistency between case facts and legal articles, our method addresses the critical challenges of logical instability and hallucination generation of LLMs in complex legal tasks. We utilized LLMs to translate NL into SL, perform deductive reasoning via a SAT solver and maintain contextual readability through final interpretation. Extensive experiments across various datasets demonstrate that our approach significantly outperforms general and domain-specific baselines.

## References

Qiming Bao, Alex Peng, Zhenyun Deng, Wanjun Zhong, Gael Gendron, Timothy Pistotti, Neset Tan, Nathan Young, Yang Chen, Yonghua Zhu, Paul Denny, Michael Witbrock, and Jiamou Liu. Abstract Meaning Representation-based logic-driven data augmentation for logical reasoning. In *Findings of the Association for Computational Linguistics ACL*, 2024.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2898–2904, 2020.

Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert van Rooij, Kun Zhang, and Zhouchen Lin. Empowering llms with logical reasoning: A comprehensive survey. *International Joint Conference on Artificial Intelligence, Survey Track*, 2025.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, et al. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*, 2024.

Jiaxi Cui, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian, and Li Yuan. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model, 2024.

Junyun Cui, Xiaoyu Shen, and Shaochun Wen. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *IEEE Access*, 11:102050–102071, 2023.

Chenlong Deng, Kelong Mao, Yuyao Zhang, and Zhicheng Dou. Enabling discriminative reasoning in llms for legal judgment prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 784–796, 2024.

Wentao Deng, Jiahuan Pei, Keyi Kong, Zhe Chen, Furu Wei, Yujun Li, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. Syllogistic reasoning for legal judgment analysis. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 13997–14009, 2023.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhumu Han, et al. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*, 2023.

Jiazhan Feng, Ruochen Xu, Junheng Hao, Hiteshi Sharma, Yelong Shen, Dongyan Zhao, and Weizhu Chen. Language models can be deductive solvers. In *Findings of the Association for Computational Linguistics: NAACL*, 2024.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1049–1065, 2023.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*, 2023.

Wanhong Huang, Yi Feng, Chuanyi Li, Honghan Wu, Jidong Ge, and Vincent Ng. CMDL: A large-scale Chinese multi-defendant legal judgment prediction dataset. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5895–5906, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.351. URL https://aclanthology.org/2024.findings-acl.351/.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

Cong Jiang and Xiaolei Yang. Legal syllogism prompting: Teaching large language models for legal judgment prediction. In *Proceedings of the nineteenth international conference on artificial intelligence and law*, pp. 417–421, 2023.

Fangkai Jiao, Zhiyang Teng, Bosheng Ding, Zhengyuan Liu, Nancy Chen, and Shafiq Joty. Exploring self-supervised logic-enhanced training for large language models. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024.

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

Aditya Kuppa, Nikon Rasumov-Rahe, and Marc Voses. Chain of reference prompting helps llm to think like a lawyer. In *Generative AI+ law workshop*. sn, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Tongxuan Liu, Wenjiang Xu, Weizhe Huang, Yuting Zeng, Jiaxing Wang, Xingyu Wang, Hailong Yang, and Jing Li. Logic-of-thought: Injecting logic into contexts for full reasoning in large language models. In *Proceedings of Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025.

Jieyi Long. Large language model guided tree-of-thought, 2023. URL https://arxiv.org/abs/2305.08291.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. Finding the law: Enhancing statutory article retrieval via graph neural networks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2761–2776, 2023.

Lauren Martin, Nick Whitehouse, Stephanie Yiu, Lizzie Catterson, and Rivindu Perera. Better call gpt, comparing large language models against lawyers, 2024. URL https://arxiv.org/abs/2401.16212.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. Enhancing reasoning capabilities of llms via principled synthetic logic corpus. *Advances in Neural Information Processing Systems*, 2024.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2023.

Hyun Ryu, Gyeongman Kim, Hyemin S Lee, and Eunho Yang. Divide and translate: Compositional first-order logic translation and verification for complex logical reasoning. In *The International Conference on Learning Representations*, 2025.

Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. Explaining legal concepts with augmented large language models (gpt-4). *arXiv preprint arXiv:2306.09525*, 2023.

Yuxuan Wan, Wenxuan Wang, Yiliu Yang, Youliang Yuan, Jen-tse Huang, Pinjia He, Wenxiang Jiao, and Michael Lyu. LogicAsker: Evaluating and improving the logical reasoning ability of large language models. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 2022.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. Cail2018: A large-scale legal dataset for judgment prediction, 2018. URL `https://arxiv.org/abs/1807.02478`.

Fangzhi Xu, Zhiyong Wu, Qiushi Sun, Siyu Ren, Fei Yuan, Shuai Yuan, Qika Lin, Yu Qiao, and Jun Liu. Symbol-LLM: Towards foundational symbol-centric interface for large language models. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, 2024.

Jundong Xu, Hao Fei, Meng Luo, Qian Liu, Liangming Pan, William Yang Wang, Preslav Nakov, Mong-Li Lee, and Wynne Hsu. Aristotle: Mastering logical reasoning with a logic-complete decompose-search-resolve framework. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, 2025.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 2024.

Xi Ye, Qiaochu Chen, Isil Dillig, and Greg Durrett. Satlm: Satisfiability-aided language models using declarative prompting. *Advances in Neural Information Processing Systems*, 2023.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, et al. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*, 2023.

Yifan Zhang, Yang Yuan, and Andrew Chi-Chih Yao. On the diagram of thought. *arXiv preprint arXiv:2409.10038*, 2024.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5218–5230, 2020.