0x.Yuan at SemEval-2024 Task 5: Enhancing Legal Argument Reasoning with Structured Prompts

Yu-An Lu
National Chupei High School
luyuam0@gmail.com

Hung-Yu Kao
National Cheng Kung University
hykao@mail.ncku.edu.tw

Abstract

The intersection of legal reasoning and Natural Language Processing (NLP) technologies, particularly Large Language Models (LLMs), offers groundbreaking potential for augmenting human capabilities in the legal domain. This paper presents our approach and findings from participating in SemEval-2024 Task 5, focusing on the effect of argument reasoning in civil procedures using legal reasoning prompts. We investigated the impact of structured legal reasoning methodologies, including TREACC, IRAC, IRAAC, and MIRAC, on guiding LLMs to analyze and evaluate legal arguments systematically. Our experimental setup involved crafting specific prompts based on these methodologies to instruct the LLM to dissect and scrutinize legal cases, aiming to discern the cogency of argumentative solutions within a zero-shot learning framework. The performance of our approach, as measured by F1 score and accuracy, demonstrated the efficacy of integrating structured legal reasoning into LLMs for legal analysis. The findings underscore the promise of LLMs, when equipped with legal reasoning prompts, in enhancing their ability to process and reason through complex legal texts, thus contributing to the broader application of AI in legal studies and practice.

1 Introduction

The process of reasoning in legal arguments is a crucial aspect of applying legal knowledge in real-world scenarios. Mastery of this skill enables individuals to effectively address legal issues and ascertain the legality of various cases. Recently, the field of Natural Language Processing (NLP) has seen significant advancements, leading to the growing trend of utilizing Large Language Models (LLMs) as tools to augment human capabilities. Given the extensive and often complex body of legal knowledge, which can be challenging and time-consuming for the average person to learn, LLMs present an opportunity to comprehend this information and offer valuable assistance.

In light of this, the organizers of SemEval-2024 Task 5 (Held and Habernal, 2024) have compiled a dataset from the domain of U.S. civil procedure. This dataset includes introductory materials on various cases, a set of questions, potential argumentative solutions, and labels indicating the accuracy of these solutions. This initiative provides a framework for evaluating the effectiveness of LLMs in the legal arena, thereby contributing to the development of more sophisticated and capable language processing tools for legal applications.

In this task, we explored legal reasoning prompts in Large Language Models (LLMs) (Burton, 2017). Our focus was on investigating their effectiveness in differentiating argumentative solutions in civil procedure cases. The results show that by guiding an LLM to think step-by-step like a lawyer, it significantly outperforms both Chain of Thought (CoT) (Wei et al., 2023) reasoning and direct output methods.

2 Background

2.1 Related Works

Large Language Model(LLM): LLM is a kind of machine learning model in Nature Language Processing(NLP), pre-trained on large scale of text and can generate text based on previous context (Naveed et al., 2023). Beside LLM's usage in general works such as ChatGPT, LLM had also show its impressive abilities several professional fields like finance, medical and legal(Kaddour et al., 2023), such as BloombergGPT (Wu et al., 2023), Med-PaLM (Singhal et al., 2023) and ChatLaw (Cui et al., 2023).

LLM in Legal Field: Legal defined rules of human community, helping to make order to our life. But legal field have lots of professional knowledge, making obstacles to common people. Lots of legal

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 385–390

LLM or related methods are developed to solve this problem, (Cui et al., 2023) and (Nguyen, 2023) had trained the LLM in the legal field in Chinese justice, (Savelka et al., 2023) found that GPT-4 performed great in explaining legal concepts, (Savelka, 2023) found that some LLM already has legal knowledge in itself. These findings demonstrate the ability and potential of LLM to address legal-related issues.

Legal Reasoning: Legal reasoning is a kind of reasoning approach which had been used in law school teaching (Bentley, 1994), this approach initially aims to help law school students thinking legal questions in professional structure. (Burton, 2017) had make a overview of several legal reasoning approaches, such as 'CLEO' (Claim, law, evaluation, outcome). These approaches originally only used in legal field, until (Savelka, 2023) used these approaches as prompt in LLM, and found that these approach can make LLM's perform well on legal reasoning task, inspired by their works, we will try to use these approach flexible in LLM to help check the truthiness of argument reasoning in civil procedure.

2.2 Dataset Description

The dataset, developed for SemEval-2024 Task 5, focuses on the domain of U.S. civil procedure, aiming to test legal language models on their argument reasoning capabilities. It is meticulously structured to include a variety of components such as a brief introduction to each case, specific legal inquiries, proposed arguments, and in-depth analyses, making it a comprehensive tool for evaluating the nuanced understanding of legal texts. Each record within the dataset is uniquely identified and contains fields that detail the legal question at hand, a potential answer, and an indicator of the answer's accuracy (limited to the training and development subsets). Additionally, the dataset offers rich analyses, including both a focused excerpt relevant to the given answer and a complete solution explanation, along with supplementary explanations to contextualize the question further. Below Tab1 is an example of the dataset:

Attribute	Value
id	0
question	1. Redistricting. Dziezek, who resides in the Southern District of Indiana, sues Torruella...
answer	Case Study: Dziezek vs. Torruella and Hopkins
label	0
analysis	So the remaining question is whether the Western District of Kentucky, where Torruella resides, is a proper venue...
complete analysis	DLet's see. Under §1391(b)(1), venue is proper in a district where all defendants reside. But here they don't all reside in the same district...
explanation	Venue in most federal actions is governed by 28 U.S.C. §1391(b), which provides: (b) Venue in...

Table 1: Dataset example

3 Methodology

3.1 Legal Reasoning Prompts

Upon examining the dataset, we found that a significant portion of the legal knowledge pertinent to the argumentation is encapsulated within the 'Introduction' segment of the dataset. This observation suggests that the primary function of Legal Language Models (LLMs) is to facilitate reasoning from the provided text, as opposed to generating novel legal insights. Consequently, we have curated a selection of legal reasoning methodologies that adhere to the principle of meticulous analysis of the given text, progressively leading to a well-founded conclusion. The methodologies selected for this purpose are as follows:

- **TREACC** (Topic, Rule, Explanation, Analysis, Counterarguments, Conclusion): Provides a comprehensive analytical framework that includes discussions on counterarguments, aiding in the consideration and evaluation of all relevant aspects of a case in Legal Language Models (LLM).

- **IRAC** (Issue, Rule, Application, Conclusion): The fundamental structure for legal issue analysis, involving the identification of the issue, the rule, application of the rule to the facts, and drawing a conclusion.

- **IRAAC** (Issue, Rule, Application, Alternative Analysis, Conclusion): In addition to the basic steps of IRAC, this method incorporates an alternative analysis of the case, showcasing different facets of the issue.

- **MIRAC** (Material facts, Issues, Rules, Arguments, Conclusion): Emphasizes the importance of material facts and arguments by discussing them in detail before proceeding with the analysis and conclusion.

3.2 Experiments

In the devised architecture, attributes such as "question," "answer," and "explanation" were judiciously chosen to elicit from the Language Model (LLM) analyses predicated on legal reasoning, utilizing a zero-shot approach. A prompt was meticulously crafted, articulating the elements of legal reasoning methodologies, thereby casting the LLM in the role of a domain specialist tasked with the meticulous evaluation of responses in accordance with legal statutes. Moreover, the LLM was directed to encapsulate facets of the legal reasoning process within designated tags, e.g., <Topic> and </Topic>, to forestall omissions and diminish the likelihood of inaccuracies. Detailed elaboration of these prompts can be found in the appendixA for consultation.

Subsequent to the generation of analysis, these analyses were employed to instruct the LLM to adjudicate the cogency of the answers provided. The Mixtral-8x7B (Jiang et al., 2024) model, noted for its cost-efficiency and superior performance, was selected for our experimental evaluations.

4 Results

4.1 Official Evaluation Metrics

Agent	F1	Acc
TREACC	0.59	0.62
IRAC	**0.60**	**0.63**
IRAAC	**0.60**	**0.63**
MIRAC	**0.60**	**0.63**
CoT	0.53	0.58
Directly	0.49	0.55

Table 2: Performance of different methods.

As demonstrated in Table 2, various methods exhibit distinct performances on the test dataset. In contrast, methodologies such as CoT and Direct Output solely leverage "question," "answer," and "explanation" to prompt the LLM to discern the veracity of the answers.

Our analysis revealed that strategies incorporating legal reasoning prompts uniformly outperformed the CoT and Direct Output approaches, underscoring the efficacy of our methodology. Intriguingly, the IRAC, IRAAC, and MIRAC methods manifested identical performance metrics on the test dataset. A deeper examination of the prediction outcomes suggested that this phenomenon could be attributed to the pronounced similarity in the analyses engendered by these methods.

The analyses generated predominantly adhered to the stipulated procedural steps, fostering a methodical and layered approach to thinking and reasoning. With the exception of approximately 1% of instances, the data conformed to our prompts, yielding comprehensive structural outputs. An illustrative excerpt from an analysis employing the IRAC method is presented below:

> <issue>...The pivotal legal question concerns the appropriate venue within the Southern District of New York for a negligence lawsuit...</issue> <rule>...Venue is determined by 28 U.S.C. §1391(b)(3), which stipulates its appropriateness in any jurisdiction where a defendant resides, provided all defendants are state residents...</rule> <application>...In initiating a negligence suit against Lenoir and Sorrel in Canada, with Sorrel being served in New York, the query arises regarding the venue's suitability for both defendants...</application> <conclusion>...Legal ambiguities surrounding the residency requirement for defendants complicate the determination of an eligible venue for proceeding with the case...</conclusion>

Our team, identified as "Yuan_Lu" on the leaderboard, achieved a commendable 10th place among 21 participating teams.

4.2 Case Discussion

Here's a comparison of a cases that classify by **IRAC**, **CoT** and **Directly** output methods.

- *Question*: Black White Taxicab, incorporated in Kentucky, seeks to do business at the Bowling Green railroad station but faces opposition from Brown Yellow Taxicab's exclusive contract with the railroad. Given Kentucky courts' stance against such contracts as against public policy, Black White sues Brown Yellow to prevent it from enforcing the contract. Assuming diversity jurisdiction and the amount-in-controversy requirement are met, should the

federal court apply Kentucky law under the Rules of Decision Act?

- *Answer*: Apply Kentucky law under the Rules of Decision Act.

- *Explanation*: This question references the historical case of Black White Taxicab Co. v. Brown Yellow Taxicab Co., highlighting the manipulation of corporate structure to create diversity jurisdiction and bring a case to federal court where federal common law, under Swift v. Tyson, would allow enforcement of a contract deemed invalid under state law. This scenario underscores the tension between federal and state law in diversity cases prior to the Erie decision.

- *Label*: **False**

4.2.1 Directly output's answer

> *The answer is TRUE. This interpretation relies on the Rules of Decision Act as understood before the Erie decision, suggesting that the federal court would apply state law unless a federal statute, constitutional provision, or treaty dictates otherwise.*

The direct output mistakenly considers the pre-Erie understanding applicable, ignoring the significant shift in legal practice introduced by Erie Railroad Co. v. Tompkins.

4.2.2 CoT's answer

> *Answer: TRUE, with a caveat. The response suggests a nuanced approach where federal courts, under the pre-Erie doctrine of Swift v. Tyson, might not strictly apply state law but would consider it alongside a broader interpretation of common law principles, particularly regarding public policy against exclusive contracts for taxi services at railroad stations.*

CoT's analysis presents a sophisticated understanding but incorrectly aligns with the pre-Erie legal framework, not recognizing Erie's directive for federal courts to adhere strictly to state law in diversity cases.

4.2.3 IRAC's answer

> *The analysis is NOT true. Erie Railroad Co. v. Tompkins overruled the Swift v. Tyson doctrine, mandating that federal courts apply state law in diversity cases. This correct interpretation implies that, in the scenario provided, the federal court would indeed apply Kentucky law, reflecting the post-Erie legal landscape accurately.*

The IRAC method accurately reflects the current legal standard, demonstrating a correct understanding of the Erie doctrine's requirement for federal courts to apply state law in diversity cases, thus providing the most legally sound interpretation.

4.2.4 Summary

Both Direct Output and CoT inaccurately conclude based on outdated pre-Erie interpretations, with Direct Output oversimplifying and CoT inadequately applying post-Erie legal standards. In contrast, the IRAC method accurately applies the Erie doctrine, demonstrating a nuanced understanding of current legal principles by methodically breaking down the issue and applying the correct rule. This approach not only ensures precision in legal analysis but also aligns conclusions with contemporary legal frameworks, showcasing its distinct contribution to legal reasoning and highlighting the importance of structured analysis in achieving accurate legal interpretations.

5 Conclusion

This study embarked on an exploration of the synergy between Large Language Models (LLMs) and legal reasoning methodologies to enhance the processing and understanding of legal texts. By integrating structured legal reasoning prompts derived from methodologies such as TREACC, IRAC, IRAAC, and MIRAC into the framework of LLMs, we demonstrated the potential of this approach to improve the models' capacity for legal argument evaluation.

References

Duncan Bentley. 1994. Using structures to teach legal reasoning. *Legal Education Review*, 5(2).

Kelley Burton. 2017. "think like a lawyer" using a legal reasoning grid and criterion-referenced assessment rubric on irac (issue, rule, application, conclusion). *Journal of Learning Design*, 10:57–68.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases.

Lena Held and Ivan Habernal. 2024. SemEval-2024 Task 5: Argument Reasoning in Civil Procedure. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models.

Ha-Thanh Nguyen. 2023. A brief report on lawgpt 1.0: A virtual legal assistant based on gpt-3.

Jaromir Savelka. 2023. Unlocking practical applications in legal domain: Evaluation of gpt for zero-shot semantic annotation of legal texts. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, page 447–451, New York, NY, USA. Association for Computing Machinery.

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

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with 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.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance.

A Prompts of Legal Reasoning Prompts

A.1 TREACC

''' Question: question Answer: answer Explanation: explanation

Analyze the given legal case scenario following these structured steps:

<topic> Identify and briefly describe the main legal issue. </topic> <rule> State the relevant legal principles or statutes that apply to the legal issue identified. </rule> <explanation> Provide a detailed explanation of the legal principles or statutes, including their background, scope, and examples of their application in previous cases. </explanation> <analysis> Apply the facts of the case to the legal principles or statutes, and evaluate how these facts fit or support the rules. </analysis> <counterarguments> Identify and explain any potential counterarguments or opposing views to the main analysis. </counterarguments> <conclusion> Summarize the analysis and provide a clear conclusion or opinion on the main legal issue. </conclusion> Use the given data to perform a structured analysis and present your findings under each labeled section. Don't forget to add label to each part, Once you're sure all tags have been added, say "I'm sure I've added all tags" at the end. Streamline the length. '''

A.2 IRAC

''' Question: question Answer: answer Explanation: explanation

Analyze the given legal case scenario following these structured steps:

<issue>Identify the key legal issue at the heart of the scenario.</issue> <rule>Detail the specific laws or legal principles that govern the identified issue.</rule> <application>Examine how the laws or principles apply to the facts of the case, discussing the legal merits of the case based on this application.</application> <conclusion>Conclude by synthesizing the analysis to state the likely outcome of the case based on the application of the rule to the issue.</conclusion>

Use the given data to perform a structured analysis and present your findings under each labeled section. Don't forget to add label to each part, Once you're sure all tags have been added, say "I'm sure

I've added all tags" at the end. Streamline the length. "'

A.3 IRAAC

"' Question: question Answer: answer Explanation: explanation

Analyze the given legal case scenario following these structured steps:

<issue>Identify the central legal issue present in the case.</issue> <rule>Articulate the rule of law that applies to the issue, including any relevant legal standards or precedents.</rule> <application>Analyze how the rule of law should be applied to the particular facts of the case, considering all relevant factors.</application> <alternative_analysis>Discuss an alternative legal analysis or perspective that might lead to a different outcome, considering other possible interpretations of the law or facts.</alternative_analysis> <conclusion>Provide a final conclusion that takes into account both the primary and alternative analyses, and state the most persuasive legal position.</conclusion>

Use the given data to perform a structured analysis and present your findings under each labeled section. Don't forget to add label to each part, Once you're sure all tags have been added, say "I'm sure I've added all tags" at the end. Streamline the length. "'

A.4 MIRAC

"' Question: question Answer: answer Explanation: explanation

Analyze the given legal case scenario following these structured steps: <material_facts>Begin by presenting the material facts of the case, focusing on those critical to the legal issues.</material_facts> <issues>Identify the specific legal issues that arise from these material facts.</issues> <rules>State the legal rules and principles that will be used to address these issues.</rules> <arguments>Develop arguments that apply the legal rules to the issues, considering the material facts and any relevant legal arguments, including policy considerations where applicable.</arguments> <conclusion>Conclude with a summary that encapsulates the findings from the application of the rules to the issues, supported by the arguments, and clearly state the resolved position on the case.</conclusion>

Use the given data to perform a structured analysis and present your findings under each labeled section. Don't forget to add label to each part, Once you're sure all tags have been added, say "I'm sure I've added all tags" at the end. Streamline the length. "'