Chain Of Reference prompting helps LLM to think like a lawyer

Aditya Kuppa [*1] **Nikon Rasumov-Rahe** [*1 2] **Marc Voses** [*3]

Abstract

Legal professionals answer legal questions based on established reasoning frameworks e.g. Issue, Rule, Rule, Application, Conclusion (IRREAC). We propose a novel technique named chain of reference (CoR) where legal questions are pre-prompted with legal frameworks thus decomposing the legal task into simple steps. We find that large language models like GPT-3 improve Zero-Shot performance by up to 12% when using the chain of reference.

1. Introduction

Legal professionals must utilize their analytical and reasoning abilities to navigate various laws and apply them to novel scenarios (Abdallah et al., 2023). The complexity and ambiguity of legal texts, added to the non-obvious nature of well-reasoned justifications, often result in inconsistent applications across different situations. However, attorneys effortlessly undertake these tasks and adeptly reason/explain the underlying rationale of their work in any language (Hoppe et al., 2021; Hoshino et al., 2019; Askari et al., 2022).

The emergence of advanced language models has demonstrated their impressive performance in diverse tasks, including zero-shot and few-shot scenarios. Nevertheless, providing satisfactory answers to legal questions necessitates more than just accurate responses. It requires a comprehensive explanation supported by a **chain of references** to relevant statutes or a legal reasoning technique, as outlined in Table 2.

In recent years, there has been a significant emphasis on utilizing reasoning-based prompt approaches to enhance the performance of large language models in question-answering tasks. To solve complex reasoning tasks using LLMs, (Wei et al., 2023) few-shot chain-of-thought (CoT) prompting was proposed that enables models to generate intermediate reasoning steps before predicting the final answer with a few shot examples. Prefixing trigger sentences such as with "*Let's think step by step*" to the prompt makes the reasoning capability more inherent. LLM can perform similarly to few-shot CoT without in-context examples.

Although Zero-shot-CoT (ZS-CoT) has successfully tackled multi-step reasoning tasks, its performance on some complex reasoning tasks poses challenges, such as semantic misunderstandings, missing some steps, etc (Li, 2023). To address the limitations of ZS-CoT and better handle the intricate nature of legal discourse, we propose a novel approach called **Chain of Reference (CoR)** Prompting. This approach comprises two key components. Firstly, we partition the given legal text into segments corresponding to different parts of the legal reasoning framework, which we refer to as references. Secondly, we apply the specific task at hand to the segmented text, enabling the language model to grasp the complete context of the legal task. By leveraging this CoR Prompting technique, we aim to enhance the understanding and performance of language models in legal tasks. Figure 1 illustrates the prompt chain for CoR. Despite the straightforward Chain of Reference (CoR) strategy, it significantly enhances the quality of the generated reasoning process. Furthermore, this prompting strategy can be easily adapted to address other legal reasoning tasks. The versatility of the CoR strategy allows for its application in various contexts, expanding its potential to improve reasoning capabilities across different problem-solving scenarios in legal contexts.

We assess the effectiveness of our proposed prompting approach using the dataset from the University of Alberta's annual Competition on Legal Information Extraction/Entailment (COLIEE) event (Rosa et al., 2022). The COLIEE dataset includes specific subtasks that involve reasoning through legal hypotheses based on contextual articles. This dataset requires providing a yes/no response and supporting legal statutes to validate the proposed hypothesis. This evaluation provides valuable insights into the performance and applicability of our prompting approach within the legal domain.

[*]Equal contribution [1]Maxime Tools, Inc, San Francisco, CA, USA [2]Meta, Menlo Park, CA, USA [3]Clyde Co, The Chrysler Building 405 Lexington Avenue 16th Floor, New York, NY 10174. Correspondence to: Nikon Rasumov-Rahe <nikon@maxime.tools>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

```
COR Prompt Stage 0:
Given the user input {{input text }}

Can you think step by step and break down
the input into chain of References?

[Issue]
[Thesis/Claim]
[Rule/Law]
[Evaluation/Evidence/Principle]
[Application/Apply]
[Reasoning/Evaluation]
[Conclusion/Outcome/Policy]


Task Prompt Stage 1:
Using the COR and user Input
COR:  {{COR TEXT}}
can you {{task definition }}
Input: {{user input }}

Output:
```

Figure 1. CoR Prompt Template. Stage 0 conditions the model to add the input text's legal reasoning steps. Stage 1 is the actual task of the LLM has to perform for example clause classification, legal QA etc

2. Experiments

In our experiments, we utilized GPT-3 as the language model (LLM) and evaluated its performance on the COLIEE 2021 test sets. Specifically, we employed OpenAI's GPT-3 text-davinci-003 variant for our experiments. To evaluate the performance of our models, we employed accuracy as the evaluation metric. The COLIEE test sets were designed with an approximately equal distribution of positive and negative answers, making accuracy a suitable metric for our evaluation. In all experiments, we set the temperature parameter to 0, corresponding to no randomness in the generated output. To maintain consistency with a temperature of 0.0, we set the top_p parameter to 1, the frequency_penalty to 0, allowing repetition by not applying any penalties to repeated tokens and the presence_penalty to 0, ensuring that no penalties were applied to tokens appearing multiple times in the output.

3. Results

Our Zero-SHOT Chain of Reference Prompting results outperform the current state-of-the-art methods by large margins as summarized in Table 1. Our experiments' results demonstrate varying accuracy levels based on the number of shots used. In the 1-shot scenario, the accuracy achieved is 0.7160. Comparatively, the zero-shot approach achieves an accuracy of 0.7407. When the number of shots is increased to 3 or 8, the accuracy improves by 2.53% for 1-shot and 1.77% for zero-shot. However, when incorporating CoT method, the accuracy decreases by 15.11%. On the other hand, employing the CoR method significantly improves accuracy, reaching 0.8348; this represents an impressive increase of 12.41% compared to the zero-shot approaches. These results highlight the impact of different shot settings and the effectiveness of incorporating legal reasoning prompts for improved performance.

Table 1. Accuracy of GPT-3's performance on the 2021 COLIEE test sets

NUMBER OF SHOTS	ACCURACY
1-SHOT	0.7160
3-SHOT	0.7531
8-SHOT	0.7531
ZERO-SHOT	0.7407
ZERO-SHOT COT	0.6296
ZERO-SHOT COR	0.8348

4. Conclusion

We propose a novel Chain of Reference Prompting technique that embeds established legal reasoning frameworks as valuable context and guidance to the language model during the generation process. This integration allows the model to generate responses that align more closely with legal norms and reasoning, leading to improved performance on established evaluation metrics. By incorporating these techniques into the prompt design, we can guide the language model toward generating more accurate, contextually appropriate, and legally sound responses.

Furthermore, previous research indicated that certain domains like legal need structured/rule-based coding data set to improve accuracy. Our findings indicate that this might be an artifact of using chain-of-thought techniques derived from coding. Using native rule-based frameworks inherent to the legal domain at question improves the accuracy without needing external coding data sets.

We strongly think the CoR technique can be leveraged in other domains, such as medicine and cybersecurity, where there are established reasoning frameworks to guide the LLMs. By leveraging the CoR technique in these domains, LLMs can harness the power of thought references to improve their performance and provide valuable insights and recommendations. This approach can enhance LLMs' accuracy, reliability, and efficiency in these critical domains, supporting professionals and decision-makers in making informed and effective decisions.

References

Abdallah, A., Piryani, B., and Jatowt, A. Exploring the state of the art in legal qa systems, 2023.

Askari, A., Verberne, S., and Pasi, G. Expert finding in legal community question answering. In Hagen, M., Verberne, S., Macdonald, C., Seifert, C., Balog, K., Nørvåg, K., and Setty, V. (eds.), *Advances in Information Retrieval*, pp. 22–30, Cham, 2022. Springer International Publishing. ISBN 978-3-030-99739-7.

Hoppe, C., Pelkmann, D., Migenda, N., Hötte, D., and Schenck, W. Towards intelligent legal advisors for document retrieval and question-answering in german legal documents. In *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pp. 29–32, 2021. doi: 10.1109/AIKE52691.2021.00011.

Hoshino, R., Taniguchi, R., Kiyota, N., and Kano, Y. Question answering system for legal bar examination using predicate argument structure. In Kojima, K., Sakamoto, M., Mineshima, K., and Satoh, K. (eds.), *New Frontiers in Artificial Intelligence*, pp. 207–220, Cham, 2019. Springer International Publishing. ISBN 978-3-030-31605-1.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, Y. A practical survey on zero-shot prompt design for in-context learning, 03 2023.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*, 2023.

OpenAI. GPT-4 technical report, 2023.

Rosa, G., Bonifacio, L., Jeronymo, V., Lotufo, R., and Nogueira, R. Billions of parameters are worth more than in-domain training data: A case study in the legal case entailment task. *Proceedings of the Sixteenth International Workshop on Juris-informatics (JURISIN 2022)*, 2022.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models, 2023.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., et al. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214–229, 2022.

5. Appendix

Table 2. VARIOUS LEGAL REASONING APPROACHES

APPROACH	DETAILS
TRRAC	THESIS, RULE, RULE, APPLICATION, CONCLUSION
CLEO	CLAIM, LAW, EVALUATION, OUTCOME
ILAC	ISSUE, LAW, APPLICATION, CONCLUSION
IRAACP	ISSUE, RULE, APPLY, APPLY, CONCLUSION, POLICY
IRREAC	ISSUE, RULE, RULE, APPLICATION, CONCLUSION
IGPAC	ISSUE, GENERAL RULE, PRECEDENT, APPLICATION, CONCLUSION
IPAAC	ISSUE, PRINCIPLE, AUTHORITY, APPLICATION, CONCLUSION
IRRAC	ISSUE, RULE, REASONING, APPLICATION, CONCLUSION
IRAC	ISSUE, RULE, APPLICATION, CONCLUSION

Hallucinations One of the prominent challenges of using LLM in the legal domain is hallucination (Ji et al., 2023), where the model generates responses that contain incorrect or false information (Weidinger et al., 2021; 2022). Despite these inaccuracies, the model can still produce seemingly coherent answers (Mahowald et al., 2023). The recent release of GPT-4 reflects this limitation, as its authors acknowledge its lack of full reliability (OpenAI, 2023). One way to mitigate the problem is to provide explicit contextual constraints augmented by a grounded knowledge base in the prompt that can guide the language model toward generating more accurate and contextually appropriate responses. The likelihood of hallucinations can be reduced by constraining the model's output to align with known facts or specific guidelines. In Chain of Reference prompting, we can reference specific sources or facts within the prompt. This can be achieved by citing reliable information from reputable sources or leveraging domain-specific knowledge bases. By explicitly referencing these sources, the language model is reminded of the relevant information it should consider when generating a response.