Structured Legal Argumentation with LLMs: A Study in Landlord-Tenant Law

Gregoire FOURNIER [a] and Daniel W. LINNA [b]
[a] *Department of Mathematics, Statistics, and Computer Science, University of Illinois Chicago, USA*
[b] *Pritzker School of Law & McCormick School of Engineering, Northwestern University, USA*

Abstract. Large Language Models (LLMs) have demonstrated capabilities across a wide range of tasks, including natural language understanding, language generation and generating basic reasoning. To improve the accuracy and reliability of these models, various prompting, augmentation, and fine-tuning strategies have been employed. For example, Chain-of-Thought (CoT) instructions have shown to be useful to guide a model's ability to solve problems. In this paper, we use context augmentation and CoT instructions to generate legal arguments for specific landlord-tenant problems using OpenAI's ChatGPT-4o. We tested this approach using ten hypothetical landlord-tenant scenarios, five provided by a legal aid organization, four generated with Anthropic's Claude, and one crafted by us. We evaluate each GPT-generated argument for accuracy, factuality, comprehensiveness, and relevance to the landlord-tenant problem scenario. This method of generating legal reasoning with LLMs offers the advantage of being more transparent and thus verifiable by legal professionals, while also providing valuable assistance to laypersons in drafting documents such as demand letters, which can help expand access to justice.

Keywords. Legal argument generation, GPT-4, Explainable AI, Chain-of-thought prompting, Augmented Large Language Models, Landlord-Tenant law

1. Introduction

Large Language Models (LLMs) have contributed to significant advancements in Natural Language Processing (NLP), demonstrating improved performance in tasks such as symbolic reasoning, language generation, and knowledge utilization [1][2]. At the same time, LLMs suffer from limitations such as hallucinations and stochastic behavior, which can undermine their accuracy, factuality, and reliability [3][4].

Techniques such as Chain of Thought (CoT) instructions [5], which guide the generation by breaking complex tasks into logical intermediate steps, have been developed to address these shortcomings.

In this paper we evaluate the ability of an LLM to generate accurate, factual, relevant, and comprehensive legal reasoning through a specific formalism of generating arguments: exposition (part of the input scenario), a specific law, how this law applies to given facts, and a conclusion.

Having the LLM generate the reasoning step-by-step, rather than only the conclusion, allows a user to review and understand the output. Presenting the argument in this

way also increases the possibility that a user can assess the argument. However, the potential for non-expert users to contribute to the legal-reasoning process as the human in the loop needs to be studied further. It is essential to explore how to design AI systems so that the human user can contribute to improving the reasoning and overall effectiveness of the system. Even a human user who is not a lawyer with expertise in landlord-tenant law can contribute to improving the accuracy, factuality, robustness, and overall effectiveness of a system. Once the system has generated the arguments, the user can take the next step, such as using the output to create a demand letter requesting that the landlord comply with legal requirements (or perhaps determining that the situation requires consulting a lawyer).

2. Related Work

LLMs have been explored for their capabilities in legal contexts. The LegalBench benchmark measures the legal reasoning abilities of 20 different LLMs [6] on 162 legal tasks. LLMs have been used to extract structured representations from legal texts to support expert systems [7], predict rhetorical roles in legal cases using GPT-3.5-turbo [8], and assist in thematic legal analysis [9]. Methods of prompt engineering and contextual provision for the practical application of LLMs have been explored in the context of insolvency law [10]. Further studies examined the ability of LLMs to annotate legal texts [11], apply tax law [12], analyze court opinions for the interpretation of legal concepts [13], and provide legal information [14]. Savelka et al. (2023) investigated how GPT-4 could explain legal concepts to professionals by augmenting prompts with relevant case-law citations based on information retrieval systems [15]. In [16], Westermann defined a semi-structured legal reasoning framework based on LLMs, which consists of generating legal reasoning by applying a given reasoning template to an input.

Our approach combines elements from the latter two works by generating legal reasoning from an augmented context. We use GPT-4o [17] to generate arguments from realistic scenarios, utilizing context augmentation of Chicago's Landlord and Tenant ordinance and a specific CoT prompt to generate relevant legal arguments. We evaluate and analyze the effectiveness of this approach using the metrics of accuracy, factuality, comprehensiveness, and relevance to the landlord-tenant problem scenario.

3. Method

We test the capability of OpenAI's latest model, GPT-4o, to generate legal arguments by applying Chicago's Residential Landlord and Tenant Ordinance (RLTO) to a specific factual scenario entered by a hypothetical user. We keep the parameters at the standard values: a temperature of 0.7, top_p at 1, and frequency and presence penalties at 0. To generate legal reasoning, we ask GPT to build each legal argument around a specific section of the RLTO. Given a legal scenario, each argument proceeds through four key steps. (1) Exposition: a relevant part of the scenario is identified to structure the argument. (2) Law: a specific section of the RLTO is cited to serve as the foundation for reasoning. (3) Application: the applicability of the cited law to the facts from the exposition is explained. (4) Conclusion: steps (1)-(3) are summarized, and the conclusion reached by applying the law to the exposition provided. The prompt is shown in Figure 1.

Input Exposition: *[Provide a detailed description of the legal scenario, including relevant facts, context, and specific issues at stake.]*

Task: Generate structured legal arguments based on the exposition provided.

Tenant and Landlord Laws considered are the Residential Landlord and Tenant Ordinance of the City of Chicago provided below:

[List specific landlord-tenant law relevant to the scenario.]

Output Format:

Argument 1:
Exposition: Summarize relevant facts from the input that relate to this argument.
Specific Law: Identify a specific law or statute that applies to the scenario.
Why This Law Applies: Explain how and why this law is relevant to the facts presented.
Conclusion: State the conclusion derived from the application of this law to the facts provided.

[Continue generating arguments as necessary, each focusing on a different applicable law.]

Figure 1. The prompt in GPT-4o used to generate legal arguments. The text in bold remains unchanged, while the text in italics is replaced with, respectively, a legal scenario and the full text of Chicago's RLTO.

4. Experimental Design

4.1. Metrics and Protocol

To evaluate legal reasoning, we choose the following metrics, the two first scored from 0 to 1, and the two others scored either 0 or 1. **Accuracy** for a given legal scenario measures how closely the set of generated arguments aligns with the true or expected answer. **Comprehensiveness** measures how well one given argument coherently and concisely addresses the relevant aspects of the input legal scenario regarding the legal requirement cited. **Factuality** assesses whether an argument originates from an actual section of the RLTO. **Relevance** evaluates whether the argument logically relates to the legal scenario.

We assess our method through the ten legal scenarios summarized in Table 1. Scenario #4 takes the landlord's perspective, while the other scenarios are from the tenant's perspective. The evaluation was performed by an expert in Landlord-Tenant law.

4.2. Experiments

Our method generated 55 arguments across the ten legal scenarios. Our experiments' scenarios and results are available online[1].

Accuracy. The accuracy of the output on eight of the scenarios was scored 1, which means that for those scenarios, the reasoning of at least one of the arguments was what our expert would have employed. In those arguments, GPT-4o was able to identify the legal requirement to apply, justify why it was applicable, and reach the correct conclusion. Scenario #7, about a landlord, asking whether one committed a crime, and Scenario #9, about privacy, were scored 0 and 0.5 for accuracy, respectively. Our expert concluded that the correct law to apply for those two scenarios is not in the RLTO. Those results

[1] `https://github.com/ssggreg/Structured-Legal-Argumentation-with-LLMs`

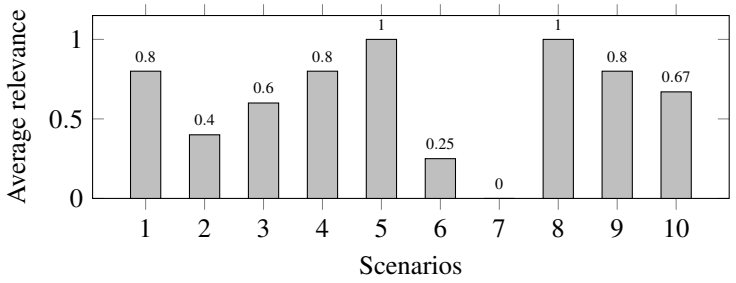

Figure 2. The average relevance of the arguments generated by GPT-4o with our method for each scenario. We distinguish two types of scenarios: #2,6,7 where most arguments were not relevant to the correct legal reasoning, and the other scenarios where most arguments dealt with the important part of the scenario.

illustrate possibilities for our approach to produce legal reasoning while also showing limitations.

Our method handled scenarios with multiple issues (scenarios #1 and #5) and was also able to successfully argue about regular wear and tear (scenario #10). We suspect that the LLM's ability to generate a well-structured argument comes from extracting the logical structure of a law upon which the argument is subsequently built. This resonates with the approach to legal reasoning in [16], where the pattern is given explicitly in the prompt of the LLM. However, when confronted with legal scenarios whose scope went beyond the RLTO, the model failed to realize its inability to complete the task.

Factuality. The factuality total score of the 55 arguments was 54. The one mistake is due to the model's poor reformulation of an RLTO section. Given the legal context augmentation and strict CoT instructions to cite the law explicitly, we anticipated a high factuality score, but this was not a given.

Table 1. Scenarios and their sources, including Anthropic's Claude [18] and LCBH[2].

Scenario Number	Description	Source
1	Heating and mold problems	Anthropic's Claude
2	Rent increase	Anthropic's Claude
3	Privacy concerns	Anthropic's Claude
4	Property damage	Anthropic's Claude
5	Rentervention 1 - cockroach infestation	LCBH
6	Rentervention 2 - eviction notice	LCBH
7	Rentervention 3 - landlord asking about crime	LCBH
8	Rentervention 4 - roof leaking	LCBH
9	Rentervention 5 - landlord taking photos inside	LCBH
10	Wear and tear	Authors

Relevance. The relevance across the arguments for each scenario is presented in Figure 2. The scenarios where most arguments are relevant correspond to cases in which multiple laws are equally applicable to the facts. On the opposite, scenarios #2 and #6 problem's are respectively about a rent increase and an eviction process, both of which are specific to one or two sections of the RLTO, which explains why selecting any other argument, even if remotely applicable, does not make a lot of sense. This highlights

[2]Law Center for Better Housing

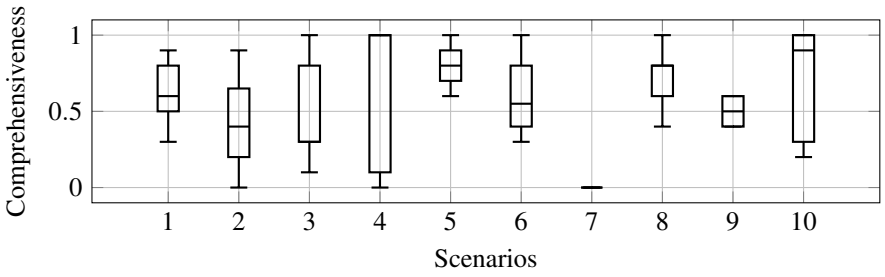

Figure 3. Boxplot representing the comprehensiveness of the arguments for the ten scenarios. In nearly every scenario, except for scenarios #7 and #9, the model successfully provided the correct legal reasoning in one of its arguments. The model, however, did not perform well at generating only high-quality arguments.

another limitation of our method for determining the strongest reasoning. We did not require the model to assert the strength of each argument, which contrasted with the expert's perspective of emphasizing a single argument whenever possible.

Comprehensiveness. The comprehensiveness of the arguments is summarized in Figure 3. First, we note that for eight out of the ten scenarios, at least one argument gave the correct legal reasoning that applies to the issue. We also see that in all but one scenario, an argument received a score inferior to or equal to 0.4. Overall, the quality of the arguments seems good, with a median greater than 0.6, but with a noticeable spread in the quality scores. This spread is more pronounced in scenarios #1-4 and #10. These scenarios were verbose, conveyed more feelings, and were written in the first person. One possible explanation is that the expert could precisely identify and focus on the relevant legal elements, efficiently filtering out irrelevant information. In contrast, the LLM's output was not focused on the relevant details.

A surprising result is that in eight out of the nine scenarios from the tenant's perspective, the model generated an argument about the landlord engaging in retaliatory actions despite this not being warranted by the scenario. This issue might be caused by a bias in the model's training data or by the model incorrectly assuming that tenants faced with one of the legal scenarios will resort to legal action.

5. Conclusion and Future Work

We used context augmentation and CoT instructions with GPT-4 to generate legal arguments for landlord-tenant scenarios. We demonstrated that LLMs with these features have the potential to generate legal reasoning. We also identified limitations, such as difficulties in classifying legal issues that fall outside the provided context and assessing the relevance of the generated arguments. Future directions include testing this framework — LLM combined with context augmentation and CoT instructions— for other tasks and with more capable reasoning models[3]. It is also valuable to compare this approach to the semi-structured legal reasoning proposed in [16]. One hypothesis is that relevant legal reasoning templates used in semi-structured legal reasoning can be extracted from specific sections of the law without additional examples.

[3]OpenAI `https://openai.com/index/learning-to-reason-with-llms/`

Acknowledgments. We would like to thank Conor Malloy from the Law Center for Better Housing for his help in providing legal scenarios and evaluating our experiments. Rentervention [4] is a free service assisting Chicago tenants with housing-related issues.

References

[1] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023.

[2] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

[3] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *ICLR*, 2020.

[4] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 2023.

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

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

[7] Samyar Janatian, Hannes Westermann, Jinzhe Tan, Jaromir Savelka, and Karim Benyekhlef. From text to structure: Using large language models to support the development of legal expert systems. In *Legal Knowledge and Information Systems*. IOS Press, 2023.

[8] Anas Belfathi, Nicolas Hernandez, and Laura Monceaux. Harnessing gpt-3.5-turbo for rhetorical role prediction in legal cases. In *Legal Knowledge and Information Systems*, 2023.

[9] Jakub Drápal, Hannes Westermann, Jaromir Savelka, et al. Using large language models to support thematic analysis in empirical legal studies. In *Legal Knowledge and Information Systems*, 2023.

[10] Marton Ribary, Paul Krause, Miklos Orban, Eugenio Vaccari, and Thomas Wood. Prompt engineering and provision of context in domain specific use of gpt. In *Legal Knowledge and Information Systems*. IOS Press, 2023.

[11] Jaromir Savelka and Kevin Ashley. The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. *Frontiers in Artificial Intelligence*, 2023.

[12] John J Nay, David Karamardian, Sarah B Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H Choi, and Jungo Kasai. Large language models as tax attorneys: A case study in legal capabilities emergence. *CoRR*, 2023.

[13] Jaromír Savelka, Kevin D. Ashley, Morgan A. Gray, Hannes Westermann, and Huihui Xu. Can GPT-4 support analysis of textual data in tasks requiring highly specialized domain expertise? In *ICAIL*, 2023.

[14] Jinzhe Tan, Hannes Westermann, and Karim Benyekhlef. Chatgpt as an artificial lawyer? In *ICAIL*, 2023.

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

[16] Hannes Westermann. Dallma: Semi-structured legal reasoning and drafting with large language models. In *GenLaw*, 2024.

[17] OpenAI. Gpt-4 technical report, 2024.

[18] Anthropic. Claude. `https://www.anthropic.com`, 2024. Large language model.

[4] `https://rentervention.com/`