If You Give an LLM a Legal Practice Guide

Anonymous Authors[1]

Abstract

Large language models struggle to answer legal questions that require applying detailed, jurisdiction-specific legal rules. Lawyers also find these kinds of questions difficult to answer. For help, lawyers turn to legal practice guides: expert-written how-to manuals for practicing a type of law in a particular jurisdiction. Might large language models also benefit from consulting these practice guides? This article examines how providing LLMs with information from legal practice guides can affect their performance with answering legal questions and predicting case outcomes. Initial findings suggest that injecting relevant excerpts from practice guides into prompts for LLMs has a negligible effect on performance. But if a practice guide is used to structure a series of LLM queries that analyze discrete issues which are then combined, LLM performance can be substantially improved. In turn, these methods can be used to comparatively evaluate legal practice guides themselves. These findings have implications for the potential for generative A.I. to automate legal tasks, particularly through agentic systems and retrieval augmented generation (RAG).

Extended Abstract

Despite being trained on vast corpuses of data, LLMs often struggle to correctly answer questions that depend upon knowledge of domain-specific information. Retrieval-augmented generation (RAG) has emerged as a method for improving LLM performance by grounding LLM responses to a set of information. (Lewis et al., 2020). With RAG, information is retrieved from a knowledge database and then injected into part of the prompt given to an LLM. RAG pipelines function like an open-book exam, giving the model a chance to answer a question using provided information.

RAG should help LLMs answer legal questions (Ajmi, 2024). One noted shortcoming of LLM performance with legal reasoning tasks is a lack of knowledge about jurisdiction-specific rules and precedent (Magesh et al., 2024). With American law, all fifty states have their own, independent state constitutions, statutes, rules of procedure, and governing caselaw. Although the current generation of LLMs have been trained on enormous datasets, LLMs still lack the granular information necessary to properly answer jurisdiction-specific legal questions (Dahl et al., 2024). RAG pipelines should be able to provide LLMs with up-to-date, jurisdiction-specific legal information.

Whether LLMs can effectively use that information is another question (Magesh et al., 2024). RAG pipelines are most successful when retrieving factual information that an LLM can directly incorporate into a response. In a legal context, RAG pipelines don't retrieve factual information but rather retrieve legal principles and rules that an LLM would then need to extract and apply. Legal questions are often complex with multiple related parts and conditional logic. Even if an application retrieves the correct legal rules, the scope or complexity of those rules may overwhelm an LLM's capacity to competently answer a question (Chen et al., 2024).

This article examines how providing LLMs with information from legal practice guides can affect their performance with answering legal questions and predicting case outcomes. Legal practice guides are a type of legal reference that helps attorneys become acclimated to a legal practice area without having to build that understanding from scratch by reading troves of statutes, regulations, and legal opinions (Davis). There are thousands of legal practice guides written on virtually every area of law for every jurisdiction in the United States. Practice guides are an ideal use case for evaluating RAG's potential for improving LLM performance at answering legal questions. Compared to other legal documents, practice guides are clearer and more succinct, and they are already formatted as instructions for attorneys to follow. With other legal documents, the verbosity and complexity of the retrieved material might cause the LLM to provide erroneous responses. Practice guides provide as clean as possible of an opportunity to observe how LLMs can apply

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

a retrieved legal rule or principle in practical scenarios.

Datasets. For datasets, we chose two practice guides covering different areas of law from different U.S. jurisdictions: a California civil practice guide for tort law and a Minnesota practice guide for real estate law. Within each guide, we selected a particular topic to test: for California, the tort law doctrine of res ipsa loquitur and for Minnesota, state law governing eminent domain. California state law about res ipsa loquitur is an example of a legal rule that is common across the United States with minimal variation. Minnesota's eminent domain laws include state-specific, unique legal rules. In future work on this paper, we will expand the dataset to be more fully representative by including practice guides that cover other areas of civil and criminal litigation in other U.S. jurisdictions.

For each topic, we excerpted from the practice guides instructions for answering a legal question on the relevant topic. These excerpts simulate a best case scenario for information retrieval, allowing us to measure the model's performance given that the correct part of a practice guide has been included within the prompt. To create data for testing LLM performance, we manually extracted the facts and holding of each relevant case referenced within that part of the practice guide. For each case, the facts provide background information on the parties and the legal dispute, and the holding provides the correct legal answer along with the reasoning behind that holding.

Methods. We used four different prompt templates to evaluate LLM performance at answering legal questions. The first two prompt templates (*Name*) and (*Fact*) served as controls to establish the LLM's baseline performance absent any help from the practice guide. (*Name*) provided the LLM with the name of the legal case, without any facts of the case or information from the practice guide. (*Fact*) provided the LLM with only the facts of the case, without information from the practice guide. The third prompt template (+*Guide*) provided the LLM with the facts of the case and the excerpt from the practice guide. The final prompt template (*Prop.*) broke the excerpt of the practice guide down into distinct components based on different parts of the relevant legal rule. A separate LLM query was made for each part of the legal rule, and the LLM was asked to evaluate whether that part of the legal rule was met. Finally, the results were combined using hardcoded Python based on propositional logic. We took advantage of probabilistic LLM outputs by requesting multiple responses from the LLM for each test, and then averaging over the results.

LLM	Name	Facts	+Guide	Prop.
GPT-3.5	0.45	0.53	0.47	0.81
GPT-4	0.45	0.75	0.86	0.75
Claude Haiku	0.50	0.63	0.68	0.91
Claude Sonnet	0.41	0.67	0.68	0.75
Claude Opus	0.53	0.56	0.70	0.66

Figure 1: Accuracy with Res Ipsa Loquitur Cases

LLM	Name	Facts	+Guide	Prop.
GPT-3.5	0.22	0.5	0.22	1.0
GPT-4	0.36	0.9	0.86	0.8
Claude Haiku	0.43	0.74	0.30	0.8
Claude Sonnet	0.16	0.86	0.45	1.0
Claude Opus	0.19	0.91	0.78	0.8

Figure 2: Accuracy with Minnesota Eminent Domain Cases

Preliminary Results. Our initial findings suggest that while LLMs can predict case outcomes better than chance given the practice guide, naively adding the practice guide results in a negligible performance increase over providing just the facts of the case — if the practice guide helps at all. This suggests that the LLMs are unable to effectively utilize the information within the practice guide. But a more involved use of the practice guide substantially improves performance. Our *Prop.* method that broke the excerpt of the practice guide down into distinct components resulted in the highest performance across tasks. One surprising finding is that although stronger models were typically more accurate than weaker models under the *Facts* method, they were often less accurate than weaker models using the *Prop.* method. Our initial investigations suggest that this may be because stronger models are more likely to introduce background knowledge beyond the provided facts, which can misdirect their legal analysis.

These findings have implications for the potential for generative A.I. to automate legal tasks, particularly through agentic systems and retrieval augmented generation (Choi & Schwarcz, 2023). Standing alone, RAG is unlikely to dramatically improve LLMs ability to answer domain-specific legal questions. But when RAG is coupled with a decomposed reasoning process, LLMs' performance can improve dramatically.

References

Ajmi, A. Revolutionizing Access to Justice: The Role of AI-Powered Chatbots and Retrieval-Augmented Generation in Legal Self-Help. 2024.

Chen, J., Lin, H., Han, X., and Sun, L. Benchmarking Large Language Models in Retrieval-Augmented Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762, March 2024. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v38i16.29728.

Choi, J. H. and Schwarcz, D. B. AI Assistance in Legal Analysis: An Empirical Study. *SSRN Electronic Journal*, 2023. ISSN 1556-5068. doi: 10.2139/ssrn.4539836.

Dahl, M., Magesh, V., Suzgun, M., and Ho, D. E. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models, January 2024.

Davis, J. LibGuides: Tort Law Research Guide: Practice Guides. https://lawlibguides.usc.edu/c.php?g=687841&p=4879061.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., and Ho, D. E. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools, May 2024.