L.E.G.A.L. (Leveraging Expert Guidance for AI in Law): A RAG-Based System for Legal Document Navigation

Hui Chia[1], Geordie Z. Zhang[2], Daniel Russo-Batterham[2], Kabir Manandhar Shrestha[2], Uy Thinh Quang[2], Zeyu Wang[2], Jeannie Marie Paterson [1]

[1]Melbourne Law School, [2]Melbourne Data Analytics Platform
The University of Melbourne, Australia
Correspondence: geordie.zhang@unimelb.edu.au

Abstract

This abstract proposal presents a Retrieval-Augmented Generation (RAG) system designed to assist users in navigating legal documents by combining large language models (LLMs) such as GPT-4 and Meta LLaMA with a curated legal database. Our approach addresses two critical challenges in the legal domain: the opacity of AI-driven tools, often referred to as "black boxes," and the risk of generating hallucinated content that is not grounded in reality. By grounding responses in verifiable legal texts, our system ensures transparency and accuracy in AI-generated legal advice. We will evaluate the system using an expert-curated legal dataset, benchmarking its performance against direct LLM prompting, while advancing towards public accessibility and open-source contributions for further research in the legal domain.

1 Introduction

The application of generative AI (GenAI) in the legal field has led to an influx of AI tools designed to assist with tasks like document analysis and case law retrieval. However, many of these tools operate as "black boxes," leaving users in the dark about how decisions are made and whether the generated information is reliable. This opacity is especially concerning in legal contexts where the cost of misinformation can be substantial. Legal professionals are often hesitant to trust such AI tools, while students and non-experts may be misled by hallucinations—fabricated information that is not grounded in the input data or legal documents (Magesh et al., 2024).

Although models like GPT-4 (OpenAI, 2023) and Meta LLaMA (Touvron et al., 2023) are powerful, they can generate responses that are not always rooted in fact. This poses a serious problem in the legal domain, where accuracy and traceability are essential. AI models must not only generate useful answers but also allow users to verify the sources of those answers.

Commercial legal research service providers have been developing tools that claim to use RAG to address the hallucination problem of LLMs. However, recent work (Magesh et al., 2024) shows that these commercially available legal research tools may not be as reliable as claimed. This highlights the need for more academic work to thoroughly evaluate the potential of RAG for legal queries and its effectiveness in mitigating hallucinations.

Our project addresses these challenges by leveraging a Retrieval-Augmented Generation (RAG) architecture which have been shown to generate more specific, diverse, and factual language compared to parametric-only models, particularly for knowledge-intensive tasks (Lewis et al., 2020). This architecture restricts generated responses to information retrieved from a curated database of legal texts, providing users with clear trails to the documents used in the reasoning process. By doing so, we mitigate hallucinations and promote trust in AI-driven legal tools, while encouraging interdisciplinary collaboration between computer scientists and legal professionals to build systems that are both robust and aligned with the specific needs of the legal domain.

2 Dataset

For our RAG system, we use the Open Australian Legal Corpus, the first and only multijurisdictional corpus of Australian legislative and judicial documents. This dataset[1], created by Umar Butler, consists of over 229,000 legal texts, including statutes, regulations, bills, and court decisions from various Australian jurisdictions.

With over 1.4 billion tokens, it provides compre-

[1] https://huggingface.co/datasets/umarbutler/open-australian-legal-corpus

hensive coverage of Australian law, making it the most suitable dataset for ensuring our system retrieves authoritative and relevant legal information. Its open-source licensing further supports accessibility, making it an ideal choice for developing and evaluating AI-driven legal tools such as ours.

3 Methodology

Our system is built around a Retrieval-Augmented Generation (RAG) framework designed to assist users in querying legal documents and receiving reliable, grounded answers. The architecture leverages the LangChain[2] framework to integrate large language models (LLMs), such as GPT-4, with a curated legal document database, as shown in Figure 1. The following steps outline how the system processes user queries:

- **Query Handling and Model Selection:** Users submit legal queries, which the system routes to the appropriate large language model (LLM) such as GPT-4 or Meta LLaMA, managed through LangChain. The model is chosen based on its suitability for the specific type of query.
- **Document Retrieval:** The system accesses a curated database of legal texts, including case law and legislation, to retrieve relevant documents based on the user's query. These documents serve as the foundation for generating accurate, grounded responses.
- **Answer Generation and Evaluation:** Using the RAG architecture, the LLM generates answers restricted to the retrieved legal documents. The system performs a self-evaluation of the generated response to check for quality and potential hallucinations. If the answer is aligned with the documents, it is marked as passing.
- **Answer Delivery and Refinement:** If the generated answer passes the evaluation, it is delivered to the user with references to the source documents. In cases where the answer does not pass, the system either performs a web search to retrieve additional information or applies alternative methods to generate a reliable response.

The pipeline, depicted in Figure 1, summarizes this process.

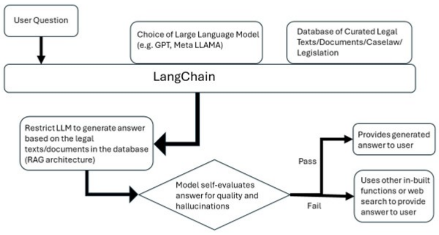
User Question
Choice of Large Language Model (e.g. GPT, Meta LLAMA)
Database of Curated Legal Texts/Documents/Caselaw/ Legislation
LangChain
Restrict LLM to generate answer based on the legal texts/documents in the database (RAG architecture)
Model self-evaluates answer for quality and hallucinations
Pass
Fail
Provides generated answer to user
Uses other in-built functions or web search to provide answer to user

Figure 1: System architecture for the RAG-based legal document navigation.

4 Evaluation

Evaluating RAG for legal queries is challenging, as multiple documents may validly answer a single query. To address this, we have created an annotated dataset of question-answer pairs along with supporting resources. Our evaluation will include BLEU score (Papineni et al., 2002) comparisons to assess alignment between RAG-generated responses and expert-provided answers, and document retrieval accuracy to measure how well retrieved documents match expert-selected sources.

We will also compare RAG performance with direct LLM prompting to assess the value of retrieval in generating grounded responses. Additionally, a crowdsourced evaluation will involve law students selecting the better answers between RAG outputs and baseline LLM outputs in a blind comparison.

5 Contributions and Future Work

The contributions of this work will be as follows:

- A legal dataset, curated with expert-verified questions, answers, and supporting references, to benchmark the system.
- A comparison between direct LLM prompting and the RAG framework in the legal domain.
- Plans to host the system for public access, with law students and legal professionals as primary users.
- The release of the source code and dataset as open-source, encouraging further research and development.

Our future work will focus on refining the system based on evaluation results, ensuring its practical application in real-world legal contexts, and expanding the dataset to cover broader legal areas.

[2] https://www.langchain.com/

References

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

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools. *ArXiv*, abs/2405.20362.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.