Self-RAG Framework for Hallucination Mitigation detected by INSIDE in Legal Analysis

Kavya Sridhar, Keisha Mehta, Kedar Desai, Yash Vishe, Guruprasad Parasnis, Marcus Chang
DSC 261 Responsible Data Science

ABSTRACT

The application of Large Language Models (LLMs) in the legal domain offers transformative potential for tasks like legal research, question answering, and case summarisation. However, the propensity of LLMs to "hallucinate", generating factually incorrect or nonexistent information, poses a critical risk in a field where precision and accuracy are paramount. Misinformation, such as citing fabricated case law or misinterpreting statutes, can have severe consequences. This project proposes a specialised Self-Retrieval-Augmented Generation (Self-RAG) system designed to address this challenge. By grounding its generation process in a trusted knowledge base of a nation's Constitution, verified case laws, and legal precedents, and by incorporating a novel self-verification mechanism, our system will ensure that all outputs are factually consistent, legally sound, and transparent. The self-verification module will critically evaluate generated text segments against retrieved evidence, identify potential hallucinations, and refine the output to produce reliable and trustworthy legal insights. This project aims to deliver a robust framework that significantly enhances the reliability of AI-assisted legal reasoning for professionals and researchers.

1 INTRODUCTION

The integration of generative AI into the legal sector is accelerating, yet its adoption is hindered by a fundamental flaw: the risk of hallucination. Standard LLMs, while fluent, lack inherent fact-checking mechanisms and can confidently present false information as truth. In the legal context, this can manifest as:

- **Fabrication of Precedents:** Citing non-existent cases or legal rulings, which can mislead legal strategy and argumentation.
- **Misinterpretation of Statutes:** Incorrectly explaining the meaning or application of laws and constitutional articles.
- **Erroneous Legal Reasoning:** Constructing logically flawed arguments based on false premises.

These errors undermine the utility and trustworthiness of AI tools for legal professionals. The high-stakes nature of legal work demands not just fluency but factual accuracy and verifiability. Existing Retrieval-Augmented Generation (RAG) systems attempt to mitigate this by providing external knowledge, but they do not inherently guarantee that the model will faithfully adhere to the provided context. There is a pressing need for a system that can not only retrieve relevant legal documents but also actively critique and verify its own generated output to ensure strict accuracy and reliability.

2 BACKGROUND AND RELATED WORK

Large Language Models (LLMs) have demonstrated impressive capabilities in legal tasks, but they often suffer from hallucinations, confidently generating plausible yet incorrect legal information [9]. Recent evaluations show that even advanced models hallucinate frequently on legal queries (e.g., producing wrong case facts or nonexistent citations in 58–82% of prompts) [9]. Such errors are not merely academic: in one notable incident, a lawyer was sanctioned after ChatGPT fabricated case law references in a brief [9]. Recognizing these risks, legal authorities have urged caution; for instance, the U.S. Chief Justice formally warned attorneys about AI-driven misinformation in his 2023 report [9]. Clearly, hallucination poses a serious challenge in the legal domain, where precision and truth are paramount [7].

To mitigate hallucinations, researchers and industry leaders are increasingly turning to Retrieval-Augmented Generation (RAG) and related techniques. RAG integrates an LLM with a trusted legal knowledge base (e.g., statutes, constitutions, case law) so that the model's outputs are directly grounded in retrieved evidence [7]. This approach has been widely promoted as a solution for domain-specific AI: leading legal research platforms like Westlaw and LexisNexis now employ RAG-based systems and claim to "avoid" hallucinations by providing hallucination-free citations [9]. Empirical results confirm that RAG can improve factual accuracy, answers tend to stay anchored to real documents rather than a model's unreliable memory [8]. For example, an evaluation of new legal GPT-powered tools found they indeed made fewer mistakes than a generic GPT-4, thanks to retrieval. However, RAG alone is not a panacea: the same study observed that even these bespoke legal AI tools still produced incorrect or mis-sourced information in 17–34% of queries [9]. A major failure mode occurs when the retriever fetches the wrong document ("document-level retrieval mismatch"), causing the model to confidently cite irrelevant or false legal text [1]. Recent research addresses this by improving the retrieval step for instance, augmenting each legal document with summaries to guide the system toward the correct source, which significantly reduces such mismatches [1].

Beyond retrieval, another promising direction is self-verification mechanisms that have the model systematically check its own output against evidence. For instance, Dhuliawala et al. (2024) introduced a "Chain-of-Verification (CoVe)" method, where the LLM drafts an answer, then generates follow-up questions to fact-check each part of that draft, answers those questions using trusted sources, and finally revises its answer, dramatically reducing factual hallucinations in open-domain tasks [6]. In the legal context, specialized systems incorporate similar self-checking. *Lawyer LLaMA*, for example, is a legal LLM that explicitly verifies statutory citations: it checks whether a cited law article actually exists and is quoted correctly, using retrieval results as a reference [10]. By catching fabricated or misquoted provisions in this way, such a system mitigates hallucinations and ensures higher reliability of its legal outputs.

Overall, the literature suggests that combining a robust retrieval-augmented pipeline with rigorous self-verification is a highly effective strategy for grounding LLMs and achieving trustworthy, hallucination-free performance in the legal domain [1, 6, 10].

3 PROJECT MANAGEMENT & COLLABORATION PLAN

Our project has four main objectives that guide both collaboration and task management:

(1) **Build a reliable legal retrieval pipeline**
- Collect and preprocess open legal datasets (e.g., Caselaw Access Project, Pile of Law).
- Index documents using FAISS or Chroma for efficient retrieval.
- Verify document provenance and maintain a data card for each source.

(2) **Develop a Self-RAG system with self-verification**
- Integrate an LLM with the retriever to generate legally grounded responses.
- Implement a self-checking loop comparing generated text with retrieved evidence.
- Log and *analyze* hallucination cases to guide iteration.

(3) **Evaluate factuality and trustworthiness**
- Use LegalBench-RAG and Bar Exam QA benchmarks.
- Measure retrieval accuracy, citation correctness, and hallucination frequency.
- Include human inspection of selected outputs to verify reasoning soundness.

(4) **Ensure ethical, transparent collaboration**
- Follow Responsible Data Science principles: use only public data, record dataset sources, and check for bias.
- Maintain transparency in communication, code versioning, and evaluation documentation.

Project Timeline

Week	Focus	Expected Outcome
1–2	Background research and dataset setup	Selected datasets + project environment ready
3–4	Retrieval pipeline implementation	Working retriever with verified indexing
5–6	Baseline RAG prototype	Model can answer legal queries using retrieved context
7	Add self-verification module	Outputs checked and revised using retrieved evidence
8	Evaluation and ethical review	Performance metrics + Responsible AI reflection
9	Finalization	Complete report, presentation, and demo notebook

4 DATASETS

Several well-established datasets are available for use in the legal domain. These datasets cover constitutional texts, verified case law, legal benchmarks, and statutory data.

- **Case Law and Precedent Datasets**
  - **Caselaw Access Project** — Provides digitized access to over 6.7 million cases covering 360 years of U.S. legal history, suitable for case retrieval and legal precedent tasks.
  - **HFforLegal/case-law (HuggingFace)** — Offers a comprehensive, multi-country dataset of legal decisions, centralized in a standardized format for efficient legal research across jurisdictions.
  - **Pile of Law** — Aggregates U.S. statutes, court opinions, administrative filings, and European and Canadian legal resources—an ideal base for legal domain model pretraining or robust retrieval.
- **Constitutional Law Datasets**
  - **Comparative Constitutions Project** — Provides annotated constitutional documents from around the world, useful for grounding reasoning in official constitutional text.
- **Legal RAG Benchmarks and Verification**
  - **LegalBench-RAG** — A benchmark dataset tailored for evaluating retrieval precision within legal RAG pipelines. It contains expert-annotated, context-linked query–answer pairs for assessing both relevance and adherence to cited evidence.
  - **Bar Exam QA and Housing Statute QA Benchmarks** — Offer real-world legal research scenarios for retrieval and reasoning tasks, testing model understanding of statutory and case-based logic.
  - **Kaggle RAG Financial & Legal Evaluation Benchmark** — Comprises retrieval-augmented test scenarios specifically for the legal domain, enabling fine-grained evaluation of factual grounding and evidence retrieval.
  - **GitHub Awesome Legal Data Repository** — Compiles links to various open, reputable legal datasets, including international sources, statutes, and case law, serving as a comprehensive index for legal data exploration.

REFERENCES

[1] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *Proceedings of ICLR*, 2024.

[2] C. Chen, K. Liu, Z. Chen, Y. Gu, Y. Wu, M. Tao, Z. Fu, and J. Ye. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. *arXiv preprint* arXiv:2402.03744, 2024.

[3] W. Su, C. Wang, Q. Ai, Y. Hu, Z. Wu, Y. Zhou, and Y. Liu. Unsupervised Real-Time Hallucination Detection based on Internal States of Large Language Models. In *Findings of ACL*, 2024.

[4] Z. Ji, D. Chen, E. Ishii, S. Cahyawijaya, Y. Bang, B. Wilie, and P. Fung. LLM Internal States Reveal Hallucination Risk Faced With a Query. *arXiv preprint* arXiv:2407.03282, 2024.

[5] M. Beigi, Y. Shen, R. Yang, Z. Lin, Q. Wang, A. Mohan, J. He, M. Jin, C. Lu, and L. Huang. InternalInspector (I^2): Robust Confidence Estimation in LLMs through Internal States. *arXiv preprint* arXiv:2406.12053, 2024.

[6] S. Dhuliawala, A. Asai, and H. Hajishirzi. Chain-of-Verification Reduces Hallucination in Large Language Models. In *Proceedings of ACL*, 2024.

[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, M. Kučerová, and V. Stoyanov. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of NeurIPS*, 2020.

[8] M. Jeong, H. Park, D. Kim, S. Cho, and S. Lee. Self-BioRAG: Improving Biomedical Reasoning through Self-Reflective Retrieval-Augmented Generation. *Bioinformatics*, 2024.

[9] Stanford Institute for Human-Centered Artificial Intelligence (HAI). Generative AI and the Law: Risks of Hallucination and Misinformation. Stanford University Report, 2023. Available at: https://hai.stanford.edu/news

[10] J. Kim, S. Oh, and H. Lee. Lawyer LLaMA: Retrieval-Augmented Legal Reasoning with Self-Verification. *Applied Sciences*, MDPI, 2024.