Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 46 No. 3 (2025)

LawRAG: Retrieval-Augmented Generation for Judicial Case Law: An Embedding Model Benchmark

Dr. L. K. Suresh Kumar[1], Mohammed Yaseen[2], Mohammed Junaid Adil[3], Dr. V. Ramesh

[1]Associate Professor, Department of Computer Science, University College of Engineering, Osmania University, Hyderabad, India

[2]Department of Computer Science, University College of Engineering, Osmania University, Hyderabad, India

[3]Department of Computer Science, University College of Engineering, Osmania University, Hyderabad, India

[4]Asst. Professor, CMRIT, Hyderabad, India

Abstract:

This paper presents LawRAG, an advanced Retrieval-Augmented Generation (RAG) system designed for legal question answering using judicial case law in the Australian legal domain. The framework integrates legal document corpora, optimized vector embeddings, and state-of-the-art large language model to produce authoritative, contextually grounded responses. Unlike prior work focused on statutory texts, LawRAG addresses the nuanced structure of court judgments through an innovative parent document retrieval strategy. This method preserves critical legal context and improves factual accuracy. We evaluate multiple embedding models on a rigorously curated legal QA dataset, identifying GTE-large as the most reliable encoder, achieving a BERT Score of 0.8476 and the highest answer relevancy (0.7444). The system's Dockerized implementation offers a fully reproducible pipeline for judicial case law analysis, establishing new best practices for contextual retrieval in legal AI applications.

Keywords: Legal AI; Retrieval-Augmented Generation; Judicial Case Law; Embedding models; Semantic Search; Question answering; Vector indexing

1. Introduction

The legal field is characterized by complex language, intricate reasoning, and a vast volume of documents, making it an ideal candidate for applications of Artificial Intelligence (AI) and Natural Language Processing (NLP). Legal professionals routinely face the challenge of retrieving precise answers from an overwhelming corpus of statutes, regulations, and judicial opinions. Traditional keyword-based research tools often fall short, lacking the ability to capture semantic nuance and contextual depth.

Retrieval-Augmented Generation (RAG) has emerged as a promising framework for legal question answering. By combining dense document retrieval with large language model (LLM) generation, RAG systems generate contextually grounded responses and reduce hallucinations, a frequent shortcoming of standalone generative models. Prior legal AI research has predominantly focused on statutory texts or regulatory documents. In contrast, our work addresses the unique challenges posed by judicial case law, which often involves layered argumentation, legal precedents, and factual specificity.

Recent efforts such as DISC-LawLLM[10] and LexDrafter[12] have demonstrated the utility of retrieval-augmented methods for legislative analysis; however, case law remains underexplored. We address this gap by introducing LawRAG, a RAG-based framework tailored for court decisions. Our contributions include a novel parent document retrieval strategy, a comparative evaluation of embedding models, and a curated dataset

for legal QA benchmarking. This paper outlines the methodology, evaluation pipeline, and practical implications of our system, aiming to support legal professionals in accessing accurate, context-aware legal knowledge.

2. Related Work

In recent years, the use of Large Language Models (LLMs) for legal tasks has grown rapidly. Researchers have explored various approaches to adapt these models for applications such as legal question answering, document drafting, and information retrieval. This has led to the development of domain-adapted legal LLMs through fine-tuning on specialized corpora. For instance,

HanFei[5] is designed for legal QA and retrieval tasks, while LawGPT_zh[6] and the broader LawGPT family [7] focus on jurisdiction-specific legal modeling using curated datasets.

In parallel, retrieval-augmented generation (RAG) has gained traction for improving factual consistency and contextual relevance in legal NLP. DISC-LawLLM[10] couples document retrieval with generative models to provide informed answers, while CBR-RAG[11] incorporates case-based reasoning to enhance legal logic and coherence. LexDrafter[12] applies retrieval mechanisms to draft legal definitions from legislative texts.

However, these systems largely emphasize statutory or regulatory texts and are frequently trained on non-English corpora, limiting their effectiveness for case law analysis. Our work addresses this gap by focusing on judicial case documents, which demand deeper semantic reasoning due to the presence of precedent, interpretive structure, and cross-referenced legal arguments.

Moreover, most existing systems retrieve short snippets, often lacking broader legal context. We apply parent document retrieval strategy to reconstruct full court cases, enabling large language models to generate more complete and accurate legal answers.

Finally, prior research rarely examines the role of sentence embedding models in RAG pipelines. We conduct a comparative evaluation of four leading encoders, GTE-large, BGE-large, MPNet, and legal-ft-1, demonstrating that embedding selection significantly affects retrieval quality and downstream generation performance.

3. Methodology

This study presents a Retrieval-Augmented Generation (RAG) system designed for legal question answering, focusing specifically on judicial case documents. The pipeline integrates document processing, vector similarity search, and large language model (LLM) generation to evaluate how different sentence embedding models affect the accuracy and relevance of legal responses.

3.1 Data Preparation and Document Indexing

We curated a dataset consisting of court decisions, primarily from the Supreme Court of New South Wales (NSW)[14]. These documents were segmented into smaller, semantically coherent text chunks to preserve legal argument flow and citation logic. In total, we generated 15,960 text embeddings for indexing. Metadata such as case IDs, jurisdiction, and citation details were preserved throughout the pipeline to ensure traceability and to support citation-based outputs.

3.2 Embedding Models and Vector Storage

Each document chunk was converted into dense vector representations using four embedding models:

- all-mpnet-base-v2
- bge-large-en-v1.5
- gte-large
- legal-ft-1

These models were selected to cover a mix of general-purpose transformers and legal-domain-specific encoders, enabling comparative evaluation across diverse semantic representations. The embeddings were stored in ChromaDB, a high-performance vector database optimized for large-scale similarity search.

3.3 Semantic Retrieval and Parent Document Expansion

Upon receiving a user query, the system encodes it using the same embedding model used during indexing. A top-*k* similarity search retrieves the five most relevant chunks. To overcome the limitations of isolated snippet retrieval, we employ a parent document retrieval strategy: case IDs linked to the top-k chunks are extracted, and all corresponding sections from those cases are aggregated. This approach reconstructs the broader legal context, allowing the LLM to access the full scope of legal reasoning and precedential structure within the original judgments.

3.4 Answer Generation via Gemini LLM

The assembled context and user query are passed to Gemini 2.0, a large language model accessed through LangChain. Thanks to its extended context window, Gemini can process large volumes of legal text effectively and generate contextually faithful answers with embedded legal citations. The model's ability to handle multi-paragraph legal input improves semantic coherence and response completeness.

3.5 Evaluation Dataset and Performance Metrics

To assess the system's performance, we developed a custom legal evaluation dataset reflecting realistic legal information needs. Using this dataset, we compared the effectiveness of the embedding models in terms of retrieval quality and factual accuracy of generated responses.

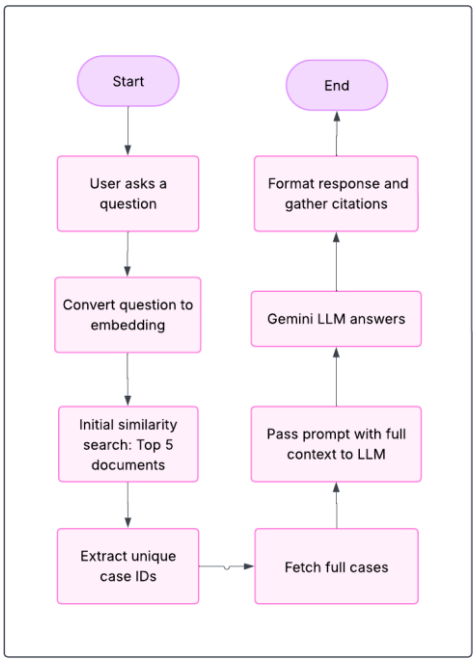

FIGURE 1. Retrieval-Augmented Generation (RAG) Workflow for Legal QA

4. Experimental setup

We developed and tested our legal question answering system using a modular RAG architecture. This section explains the technical setup, from system deployment to the evaluation process.

4.1 Containerization and Deployment

To ensure a consistent and reproducible environment, we deployed the entire RAG pipeline using Docker. Containerization provided several advantages:

- It guaranteed that the system behaved the same way across different machines and setups.
- It allowed easy scaling of components, such as the retriever or indexer, to handle larger datasets or higher user demand.
- Each module, such as indexing or retrieval, runs in its own container. This modular design makes maintenance easier, as individual components can be updated independently without affecting the rest of the system.

4.2 Document Indexing Pipeline

The indexing module preprocesses the legal corpus to enable efficient semantic retrieval. The steps involved include:

1. Document Ingestion: Legal documents are compiled from a range of publicly available judicial and legislative records.

2. Text Chunking: Each case is segmented into smaller, semantically coherent text blocks (e.g., paragraphs or logical units). Chunking at this granularity ensures that retrieval is sensitive to specific legal concepts, arguments, or facts, allowing the system to compare queries against fine-grained content rather than full-length documents. This improves retrieval precision, especially for complex or narrow legal questions.

3. Embedding Generation: Each chunk is transformed into a vector using models such as MPNet-based encoder, BGE-large, GTE-large and Legal-ft-1.

4. Vector Persistence: The resulting embeddings are stored in a ChromaDB vector store, indexed by their chunk metadata for fast similarity queries.

4.3 Retrieval Pipeline and Context Assembly

The retrieval module identifies the most contextually relevant chunks corresponding to user queries. Its operations include:

1. Query Embedding: Incoming queries are encoded into vector representations using the same model used during indexing.

2. Semantic Search: Vector similarity (e.g., cosine similarity) is computed between the query and stored embeddings to rank document relevance.

3. Top-K Initial Retrieval: The top five most similar text chunks are retrieved as the initial context. These chunks typically correspond to isolated paragraphs or segments of court cases.

4. Parent Document Expansion: While small text chunks are ideal for semantic retrieval, offering precise and fine-grained matches to user queries, they often lack the broader legal context necessary for accurate legal reasoning and generation. To address this, the system performs parent document expansion after the initial retrieval phase. The system extracts the original case id metadata associated with each of the top-5 chunks. It then retrieves all other chunks from the same case(s), effectively reconstructing the full document(s) from which

the relevant snippets originated. This allows the LLM to consider broader legal arguments and context beyond the isolated paragraphs.

5. Context Assembly: All collected chunks from the identified cases are concatenated and structured to form the input context for the language model. This ensures that answer generation remains both relevant and faithful to source material.

4.4 Large Language Model Integration

The selected context is passed to Gemini 2.0 Flash, a state-of-the-art large language model, via the Google API. The model is prompted to generate clear, accurate answers grounded in the legal content provided. Owing to its extended context window, Gemini 2.0 can process a substantial volume of retrieved information effectively, improving legal answer fidelity.

4.5 Evaluation Methodology

The system's performance was assessed using a hybrid of automated evaluation frameworks and semantic scoring:

1. RAGAS: The RAGAS framework evaluates both retrieval quality and answer generation via multiple metrics:

i. Faithfulness: Alignment of generated content with source documents.

$$F = \frac{|V|}{|S|} \tag{1}$$

where |V| is the number of statements that were supported according to the LLM and |S| is the total number of statements[18].

ii. Answer Relevance: We say that the answer as q is relevant if it directly addresses the question in an appropriate way. In particular, our assessment of answer relevance does not take into account factuality, but penalizes cases where the answer is incomplete or where it contains redundant information. To estimate answer relevance, for the given answer as q, we prompt the LLM to generate n potential questions qi based on as q, We then obtain embeddings for all questions using various embedding models. For each qi, we calculate the similarity sim(q, qi) with the original question q, as the cosine between the corresponding embeddings[18]. The answer relevance score, AR, for question q is then computed as:

$$AR = \frac{1}{n}\sum_{i=1}^{n} sim(q, qi) \tag{2}$$

iii. Context Relevance: The context c(q) is considered relevant to the extent that it exclusively contains information that is needed to answer the question. In particular, this metric aims to penalize the inclusion of redundant information[18]. To estimate context relevance, given a question q and its context c(q), the LLM extracts a subset of sentences, Sext, from c(q) that are crucial to answer q,

$$\text{CR} = \frac{\text{number of extracted sentences}}{\text{total number of sentences in } c(q)} \tag{3}$$

2. BERT Score: We use BERT Score to evaluate the semantic similarity between the generated answer and a reference (ground truth) answer. Unlike traditional lexical overlap metrics, BERT Score leverages contextual

embeddings from pre-trained language models to assess meaning, capturing nuances in phrasing. This makes it especially suitable for legal QA, where wording may differ but meaning remains consistent[19].

TABLE 1. Evaluation data for all embedding models

Field	BGE-Large	GTE-Large	MPNet	Legal-ft-1
Question	Can a freezing order be granted against a special purpose vehicle in a construction defects case?	Can a freezing order be granted against a special purpose vehicle in a construction defects case?	Can a freezing order be granted against a special purpose vehicle in a construction defects case?	Can a freezing order be granted against a special purpose vehicle in a construction defects case?
Answer	Yes, in the case of *The Owners – Strata Plan No. 102081 v Aqualand Construction Pty Ltd* [2025] NSWSC 31, a freezing order was granted against the developer, which was a …	Yes, in *The Owners \u2013 Strata Plan No. 102081 v Aqualand Construction Pty Ltd [2025] NSWSC 31*, the Owners Corporation sought a freezing order against the Developer, which was a special purpose vehicle…	Yes. In *The Owners – Strata Plan No. 102081 v Aqualand Construction Pty Ltd* [2025] NSWSC 31, a freezing order was granted against a special …	Yes, a freezing order can be granted against a special purpose vehicle in a construction defects case. In The Owners \u2013 Strata Plan No. 102081 v Aqualand Construction Pty Ltd [2025] NSWSC 31, the Owners Corporation sought a freezing order against the developer…
Key Context	- UCPR r 25.11 confers a power on a court to make a freezing order... - Plaintiff owners corporation seeks freezing order against defendant developer...	-There is no shortage of verbal formulations as to the test to be applied when considering there is such a\u201cdanger\u201d… -The Owners Corporation must also show, relevantly to the matters in dispute before me today, a danger…	- Freezing order is an extraordinary remedy requiring caution... - UCPR r 25.11 permits freezing orders to avoid frustration of judgment...	-The freezing orders had been made by Wright J in August 2024and continued by … -There can be little doubt that the first defendant was aware of the freezing order.At [164] of his affidavit he states…
Ground Truth	Yes, a freezing order can be granted against a special purpose vehicle if there is evidence that it may dispose of ...	Yes, a freezing order can be granted against a special purpose vehicle if there is evidence that it may dispose of assets in a way that would render a future judgment unsatisfied...	Yes, a freezing order can be granted against a special purpose vehicle if there is evidence that it may dispose of assets in a way that would render a …	Yes, a freezing order can be granted against a special purpose vehicle if there is evidence that it may dispose of assets in a way that would render…

5. Evaluation

To assess the performance of LawRAG in legal question answering, we developed a structured evaluation framework tailored to the specific demands of judicial case retrieval and legal answer generation. The objective was to measure how effectively different embedding models supported accurate context retrieval and semantically faithful answer generation when applied to real-world legal queries.

Metrics	Embedding models			
	Bge-large-en-v1.5	All-mpnet-base-v2	Gte-large	Legal-ft-1
Answer relevancy	0.7056	0.7033	0.7444	0.7111
Context precision	0.7500	1.0000	0.9167	0.7000
Context recall	1.0000	0.0000	1.0000	1.0000
Faithfulness	0.7500	0.0000	1.0000	0.7500
Response groundedness	1.0000	0.0000	1.0000	1.0000
BERT Score	0.8641	0.8560	0.8476	0.8448

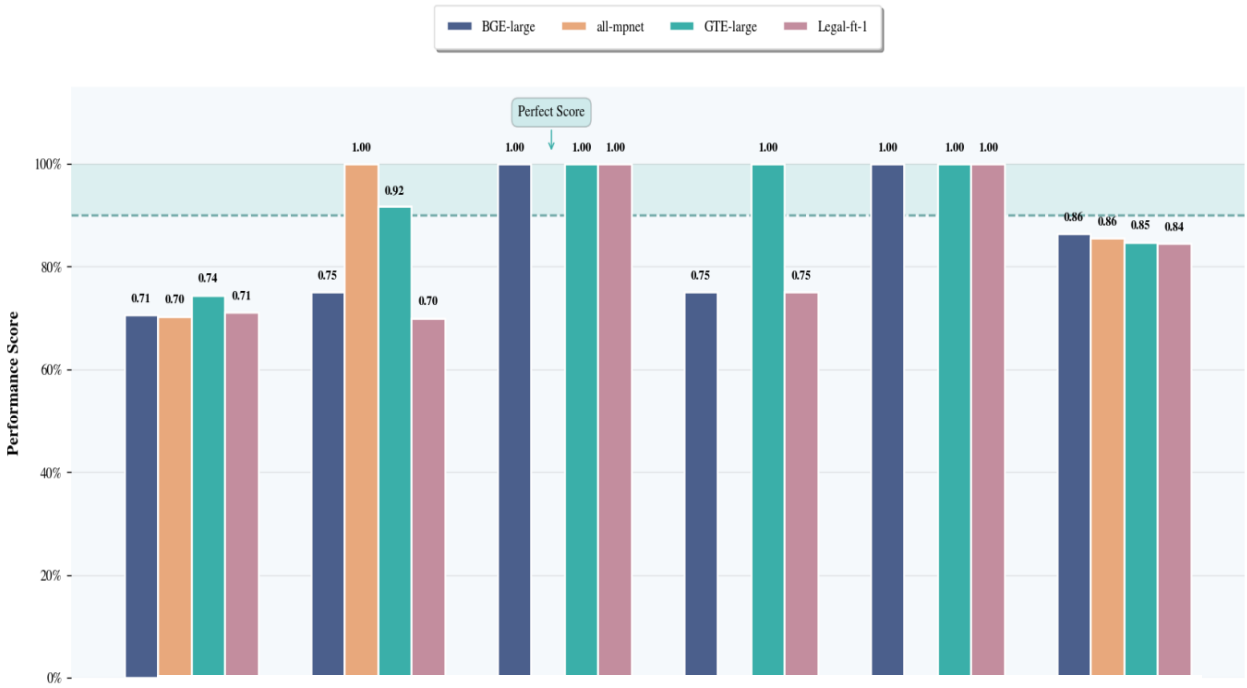

FIGURE 1. Comparison of embedding model performance across six evaluation metrics.

A custom evaluation dataset was developed, comprising 100 legal questions derived from real-world Australian court decisions. Each question was paired with three components: (i) the top-ranked context chunks retrieved by the system, (ii) the corresponding answer generated by the RAG pipeline, and (iii) a manually curated ground truth answer. The dataset was meticulously annotated by legal researchers to reflect realistic expectations for judicial question answering systems.

As illustrated in **Table 1**, we provide an example legal question alongside the retrieved contexts, system-generated responses, and reference answers for each of the four embedding models. This qualitative snapshot demonstrates how different embeddings influence the clarity, depth, and legal fidelity of generated outputs.

Table 2 summarizes the comparative performance of the four evaluated embedding models, BGE-large-en-v1.5, GTE-large, all-mpnet-base-v2, and Legal-ft-1 across six core evaluation metrics. The key findings are as follows:

- GTE-large achieved the highest Answer Relevance (0.7444) and BERT Score (0.8476), indicating strong semantic alignment with legal queries and robust answer quality.
- Legal-ft-1 demonstrated perfect Context Recall (1.000) and Response Groundedness (1.000), highlighting its ability to retrieve comprehensive case information, although its slightly lower Answer Relevance suggests room for improvement in generative precision.

TABLE 2. Performance of Embedding models on various metrics

- BGE-large-en-v1.5, performed well across all metrics, particularly in Response Groundedness (1.000) and Context Recall (1.000). However, its answers were slightly less semantically rich compared to those generated by GTE-large.

- All-mpnet-base-v2, a general-purpose embedding model, significantly underperformed, with scores of 0.000 in multiple categories. Its lack of legal domain tuning rendered it unsuitable for high-stakes legal QA.

As visualized in **Figure 1**, these results strongly underscore the importance of embedding model selection within RAG pipelines for legal applications. Models that are domain-adapted or instruction-tuned offer notable gains in factual consistency, contextual fidelity, and semantic appropriateness. In summary, the evaluation confirms that combining parent document expansion with optimized embedding models substantially improves legal question answering performance. Among all tested models, GTE-large emerged as the most balanced and effective encoder, reinforcing the conclusion that contextual grounding, model specialization, and legal-aware retrieval strategies are essential for developing accurate and reliable legal AI systems.

6. Ethical and Privacy Considerations

As legal applications of LLMs continue to expand, addressing ethical risks and privacy concerns becomes increasingly important. In RAG systems designed for legal question answering, issues such as data privacy, model bias, hallucination, and system safety require careful consideration.

One of the primary risks in this domain is bias in legal datasets. Judicial case law, while authoritative, may still reflect systemic biases or outdated legal perspectives. Embedding models and language models trained on such data can unintentionally reinforce these biases in generated answers. Additionally, models may produce hallucinated or factually incorrect responses, which can be particularly problematic in legal contexts where accuracy is critical.

Privacy is another key concern. Although our study utilizes publicly available case law, legal documents often contain sensitive information or personal identifiers. Systems that process such documents should implement safeguards to prevent accidental disclosure, especially when deployed in real-world legal settings.

Transparency and accountability are also crucial. While our RAG system generates responses based on retrieved sources, it remains essential for users to verify outputs, particularly since legal advice carries serious implications. The inclusion of source citations and traceable metadata in our system is one step towards enhancing transparency and enabling users to review the original legal context.

It is also important to recognize the role of human oversight in these systems. Although AI-powered tools can improve research efficiency, they are not substitutes for professional legal judgment. We advocate for responsible deployment of such systems, where human–AI collaboration is prioritized, and legal experts are actively involved in reviewing AI-generated outputs.

Our review of related work suggests that ethical considerations remain underexplored in many legal AI studies. Very few papers explicitly address risks related to privacy, robustness, or bias, leaving substantial room for further research in this area. We recommend that future studies integrate dedicated ethical evaluations and establish clear guidelines for responsible use in legal applications.

7. Challenges and Future Work

Legal texts, particularly judicial opinions, are structurally intricate, often embedding statutory references, layered legal reasoning, and inter-case citations. This complexity hinders accurate semantic retrieval and poses challenges for generative models to produce legally sound and contextually faithful responses. Moreover, the evaluation of legal QA remains inherently subjective, as multiple valid interpretations may exist for a given legal question, and ground truth annotations are often absent or ambiguous.

While our current system demonstrates promising capabilities for AI-assisted legal research, several avenues remain open for advancement. Future work may focus on extending the system to support multilingual corpora, thereby enhancing accessibility for diverse legal jurisdictions. Additionally, constructing a citation graph across the legal corpus could enable citation-aware retrieval and facilitate precedent ranking based on legal influence. To improve the system's adaptability, integrating a real-time ingestion pipeline for newly published judgments would ensure up-to-date information is always available for querying. Finally, embedding legal reasoning traces and argument flow visualizations could further improve system interpretability and trustworthiness.

8. Conclusions

This project introduced a Retrieval-Augmented Generation (RAG) framework specifically designed for the legal domain, aiming to improve access to precedent-based knowledge through question answering over Supreme Court case law. By combining dense retrieval methods, a Chroma-based vector store, and a large language model, the system delivers grounded, context-rich responses to user queries. The approach supports legal practitioners and researchers in identifying relevant case material efficiently and with higher semantic precision. Evaluation through RAGAS and BERT Score demonstrated that the system retrieves and generates answers that are both contextually faithful and semantically aligned.

We conducted a comparative evaluation of four embedding models BGE large, all mpnet base v2, GTE large, and Legal ft to determine the most suitable encoder for legal retrieval. Among these, GTE large consistently demonstrated strong performance across nearly all metrics. It achieved perfect context recall, faithfulness, and response groundedness, while maintaining a high BERT Score (0.8476) and superior answer relevancy (0.7444). These results suggest that GTE large effectively captures semantic intent in legal queries and retrieves precise, trustworthy contexts for LLM generation.

BGE large, while also strong, had slightly lower answer relevance (0.7056) and faithfulness, although it achieved the highest BERT Score (0.8641), indicating superior lexical-semantic matching. All mpnet base v2 achieved perfect context precision but failed in context recall and faithfulness indicating that while it retrieves highly precise chunks, it misses relevant broader context. Legal ft-1 showed balanced but slightly lower performance across all metrics.

In conclusion, GTE large emerges as the most robust embedding model for the LawRAG system. Its balance of semantic relevance, contextual accuracy, and factual consistency makes it especially well suited for retrieval-augmented legal reasoning tasks. These findings reinforce the critical role of embedding model selection in domain-specific RAG pipelines and validate the effectiveness of our system in addressing complex legal information needs.

References

[1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge Intensive NLP Tasks,"arXiv:2005.11401*, 2021.

[2] M. Park, H. Oh, E. Choi, and W. Hwang, "LRAGE: Legal Retrieval-Augmented Generation Evaluation Tool," arXiv:2504.01840v2*, 2025.

[3] A. B. Hou, O. Weller et al., "CLERC: A Dataset for U.S. Legal Case Retrieval and Retrieval-Augmented Analysis Generation," in *Proc. NAACL*, 2025, pp. 7913–7928.

[4] N. Pipitone and G. Houir Alami, "Legal Bench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain,"arXiv:2408.10343v1*, 2024.

[5] J. Wen and W. He, “HanFei-1.0,” GitHub Repository: *siat-nlp/HanFei*, 2023. Available: https://github.com/siat-nlp/HanFei

[6] H. Liu, Y. Liao, and Y. Meng, “Chinese Law Large Language Model,” GitHub Repository 2023.

[7] H.-T. Nguyen, “A Brief Report on LawGPT 1.0: A Virtual Legal Assistant,”arXiv:2302.05729*, 2023.

[8] D. Soong et al., “Improving Accuracy of GPT-3/4 Results on Biomedical Data Using a Retrieval-Augmented Language Model,” *PLOS Digital Health*, vol. 3, no. 8, 2024, Art. no. e0000568.

[9] C. Zakka et al., “Almanac: Retrieval-Augmented Language Models for Clinical Medicine,” *NEJM AI*, vol. 1, no. 2, pp. 1–45, 2024.

[10] S. Yue et al., “DISC-LawLLM: Fine-Tuning Large Language Models for Intelligent Legal Services,” arXiv:2309.11325*, 2023.

[11] N. Wiratunga, R. Abeyratne, L. Jayawardena, K. Martin, S. Massie, I. Nkisi-Orji, R. Weerasinghe, A. Liret, and B. Fleisch, “CBR-RAG: Case-Based Reasoning for Retrieval-Augmented Generation in LLMs for Legal Question Answering,” arXiv:2404.04302, 2024..

[12] A. Chouhan and M. Gertz, “LexDrafter: Terminology Drafting for Legislative Documents Using Retrieval-Augmented Generation,” in *Proc. LREC-COLING*, 2024, pp. 10448–10458.

[13] S. Alotaibi et al., “KAB: Retrieval-Augmented QA for Islamic Jurisprudence,” *IJCSNS*, 2022.

[14] *Open Australian Legal Corpus*, Available: https://openlegalcorpus.au

[15] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, “RAGAS: Retrieval-Augmented Generation Scoring Framework,” arXiv:2309.15217*, 2025.

[16] B. Saha, U. Saha, and M. Z. Malik, “QuIM-RAG: Advancing Retrieval-Augmented Generation With Inverted Question Matching for Enhanced QA Performance,” *IEEE Access*, 2025, doi:10.1109/ACCESS.2024.3513155.

[17] LLM Wizard, “legal-ft-1,” Hugging Face, 2024. Available:https://huggingface.co/llm-wizard/legal-ft-1

[18] thenlper, “gte-large,” Hugging Face, 2024. Available: https://huggingface.co/thenlper/gte-large.

[19] BAAI, “bge-large-en,” Hugging Face, 2024. Available: https://huggingface.co/BAAI/bge-large-en.

[20] Hugging Face Transformers Team, “MPNet,” HuggingFace,2024.Available:https://huggingface.co/docs/transformers/model_doc/mpnet