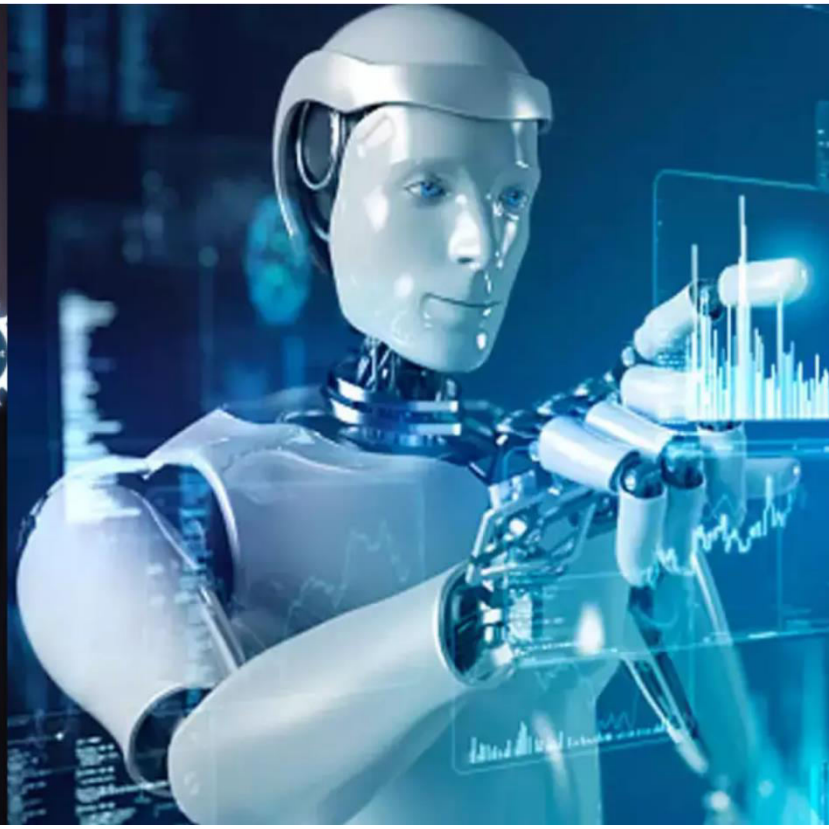
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 2, February 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502067|

LexiAI: A Retrieval-Augmented Large Language Model Framework for Legal Case Summarization and Document-Grounded Question Answering

K Shandini, V Arvind, V S Hasini, P V Pedda Reddy, N Sainath

Asst. Professor, Department of Computer Science and Engineering, Kuppam Engineering College, AP, India

Student, Department of Computer Science and Engineering, Kuppam Engineering College, AP, India

Student, Department of Computer Science and Engineering, Kuppam Engineering College, AP, India

Student, Department of Computer Science and Engineering, Kuppam Engineering College, AP, India

Student, Department of Computer Science and Engineering, Kuppam Engineering College, AP, India

ABSTRACT: Legal research entails detailed analysis of voluminous judicial data for extracting background facts, legal issues, judicial logic, and finally arriving at the conclusions of the judicial decisions. Although Large Language Models (LLMs) exhibit high-level natural language understanding capabilities, the applicability of LLMs for direct legal domain use is limited by the factors of hallucination, traceability, and the absence of context grounding.This paper presents a new model, called LexiAI, which adopts the model named Retrieval-Augmented Generation (RAG), a model suitable for structured case summarization and document-grounded legal question answering. This model combines semantic document retrieval with controlled prompt engineering, which allows it to generate strict response determination based on relevant laws. From an experimental observation, it can be noticed how the relevance of context is made better, avoiding hallucinations compared with another model.

KEYWORDS: Retrieval-Augmented Generation (RAG), Legal NLP, Large Language Models, Semantic Search, Legal Case Summarization, Vector Embeddings.

I. INTRODUCTION

The legal system is inherently document-centric. Various courts and regulatory agencies are continuously producing legal judgements, cases, statutes, and legal interpretations. It is of utmost necessity for legal professionals to peruse these documents in order to assess the background, legal issues, and arguments, as well as understanding judicial interpretations. Conventional tools for legal research currently depend on keyword-based search engines. Although these systems aid accessibility, they fail to comprehend context semantically. Legal principles are often formulated using diverse sets of terminologies; however, conventional mechanisms cannot address contextual similarities between cases. Thus, similar judicial cases may be overlooked completely.The introduction of the Transformers architecture by Vaswani et al. [1] transformed the paradigm for language modeling with contextual learning. BERT, a language model, and Word2Vec, an embedding technique, improved semantic similarity detection. However, LLMs for legal text generation are likely to generate hallucinated legal reasoning if not sourced from authoritative materials.

Furthermore, LexiAI enhances reliability and transparency in legal research by grounding every generated response in verifiable judicial sources. The system first encodes legal documents into semantic embeddings and stores them in a vector database. When a user submits a query, the model retrieves the most contextually relevant case segments before generating a response. This retrieval-first approach minimizes hallucination, improves contextual accuracy, and ensures that legal professionals receive precise, explainable, and source-backed insights. Consequently, LexiAI bridges the gap between traditional keyword-based systems and intelligent, context-aware legal research platforms.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502067|

II. LITERATURE SURVEY

Previous legal NLP tools used rule-based extraction and statistical classification techniques. LexNLP [5] offered basic tools to process legal texts using parsing techniques. Subsequent research has also investigated deep learning-based classification techniques and prediction of legal judgements [6].

Chalkidis et al. in [3] have shown the effectiveness of transformer-based models that were trained on legal data. In addition, there was Retrieval-Augmented Generation proposed by Lewis et al. in [2], which reduced hallucination in knowledge-intensive NLP tasks.

LexiAI enhances these developments with its integration of semantic retrieval as well as controlled summarization through LLM within a legal research platform.

III. PROPOSED SYSTEM

The LexiAI architecture is based on a modular and a scalable design that is able to distinguish between retrieval and reasoning functions. The system architecture is formed of a multi-stage pipeline with a distinct purpose per module. However, the importance of maintainability and extensibility cannot be overemphasized, particularly in legal domain applications.

Overall, the workflow ranges from ingesting the documents to generating output. It will be described in terms of key architectural components as follows:

• **Legal Document Ingestion:** The source of the judiciary documentation consists of publicly accessible sources like legal document repositories. These documentation entries can be judicial decisions, case laws, legal interpretations of legislature, regulations, and several other legal documentation. In the process of ingestion of the documentation entries into the system, it gets standardized into a uniform text format.

• **Preprocessing and Semantic Chunking:** Legal texts, for example, can be lengthy, and their logical unit levels may include information regarding case backgrounds, legal issues, arguments, judicial reasoning, and many other sections. This indicates a drawback of taking the entire text as a single unit, which leads to a loss of retrieval precision. As a result, the pre-processing module will be responsible for text normalization, the removal of irrelevant information contained in the formatting, and text chunking, where the entire text is divided into meaningful chunks, defining a section of the text.

• **Embedding Generation:** Each document is split into chunks, and each of the chunks is converted to a dense vector representation with the help of transformer-based language models. Unlike keyword indexes, these language models pick out concept similarity even when different terminology is used. This step converts unstructured legal documents into structured documents that are more machine-interpretable, numerical data.

• **Vector Database Indexing:** The embeddings are stored in a vector database for optimal similarity search. Vector indexing is used for the efficient query of semantically related document segments based on their distance using similarity metrics. This type of indexing technique is ideal for the scalability of large amounts of legal data and real-time processing of queries.

• **Semantic Retrieval:** Once the query is entered by the user, and the user either asks a query or requests a summary of a case, the query is embedded into a format based on the same encoding model. What is then retrieved is the semantically most relevant document chunks from the vector database. It is different from conventional search engines as this can retrieve contextually similar information even in the absence of matching keywords.

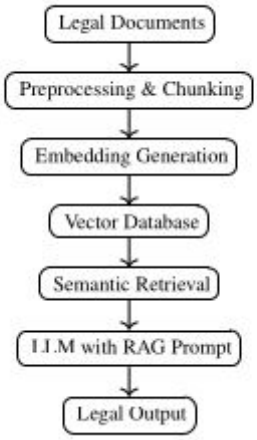
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502067|



• **Controlled LLM-Based Generation:** The last phase includes structured prompt development, as well as controlled response generation with the input of a Large Language Model. The afore-retrieved document chunks are input into pre-designed prompt templates that incorporate clear directions to ensure that response generation is entirely based on or traceable to the input context. The generated text could either be structured case summaries or document-informed responses to legal queries.

This is achieved because of the modularity of LexiAI's architecture, ensuring that the retrieval section and generative sections of the conversation flow remain decoupled. This, in turn, adds to the reliability of the system, as it does not allow for unrestricted generative reasoning while generating its response. It can also be extended with better embedding models, alternative vector databases, or fine-tuned language models without having to revisit the entire architecture.

III. METHODOLOGY

The core intelligence of the LexiAI system is powered by the Retrieval-Augmented Generation framework, which merges semantic retrieval with controlled generative modelling to ensure responses are always based on authoritative legal documents. Unlike other generative systems that usually depend on internal model knowledge, this kind of RAG approach explicitly embeds retrieved contextual evidence before generating text.

In standard deployments of Large Language Models, responses are generated by the model based on their pretraining parameters. This allows for fluency and coherence in the generated texts but can also lead to hallucinations, fabricated citations, or unsupported interpretations—especially within specialized domains such as law. Legal applications require a high degree of reliability, traceability, and factual accuracy. For these reasons, unfiltered generative reasoning is inadequate for professional legal research environments.

The RAG pipeline in LexiAI follows a structured multi-stage procedure:

1) Query Encoding

When a user submits a legal query or requests a summary of a case, the input text is first embedded into a dense semantic space utilizing the same transformer-based encoder used during document indexing to guarantee the representations are consistent between queries and stored document segments.

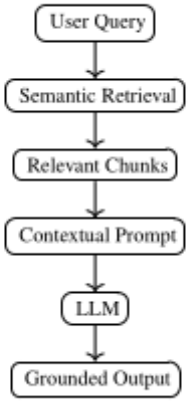
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502067|



2) Semantic Retrieval

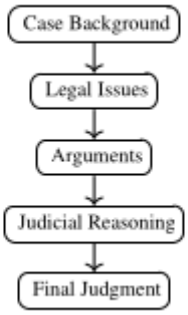
The query embedding is compared to the existing document embeddings present within the vector database. Using similarity search mechanisms, the most relevant document chunks are extracted. Unlike a keyword-based search, the process is able to discover more conceptual relationships between pieces of text, with relevance of legal reasoning able to be identified even when other terms have been used.

3) Context Construction

These retrieved segments of documents are arranged into a structured context block. There is careful ordering and formatting of retrieved documents to ensure logical flow and legal coherence. This context block serves as the evidence for the generated responses.

4) Controlled Prompt Engineering

The contextual block is placed into a specific prompt template. The explicit instructions, within the prompt, set up the language model to strongly output the retrieved content verbatim. This instruction serves as guardrails, reducing the chance of hallucination and making sure the output aligns with source documents.



5) Grounded Response Generation

The output, either structured summary format or legal answers, is produced by the Large Language Model, and it is conditioned by the supplied context. The process of generating is informed by evidence, and thus it is always traceable to the corresponding judicial text.

The hybrid approach to retrieval and generation enables several benefits. First, it increases the level of accuracy given that it ensures the model's reasoning capabilities are restricted to available information. Second, it enables increased transparency due to the ability to trace information to their raw forms. Finally, it enables increased adaptability as improvements to the embedding models are easily captured without requiring modifications to the generation component.

By combining semantic retrieval and controlled generation in a single workflow, LexiAI ensures the effective utilization of the advantages of Large Language Models while overcoming some of the limitations of the model.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502067|

Therefore, the RAG framework provides the foundation for the implementation of legal AI assistance as an effective component.

IV. EXPERIMENTAL RESULTS

In order to determine the efficacy of LexiAI, a set of public judicial judgments relating to specific circumstances was obtained from public legal repositories. The gathered data was then subjected to text normalization and semantic segmentation.Each document has been segmented into relevant sections, including background, issues, judicial reasoning, etc. The dense vector representations have been created using transformer encoders, which have then been indexed using a vector database.Evaluation criteria were qualitative and task-oriented. For the purpose of information retrieval, the relevance of the system developed was assessed based on whether or not it rendered appropriate document segments for the user queries. For the summarization system, system consistency was assessed based on its clarity, completeness, and consistency with the original content of the case.An experimental setup was developed to mimic various legal research scenarios, making the performance of the system as close to natural usability as possible.

The experimental findings show that the accuracy of context is improved through using LexiAI, compared to traditional search engines, such as keyword search systems. The semantic retrieval mechanism successfully identified segments of documents, regardless of whether they specifically contained the keyword of interest. In turn, the structured summaries provided by the system were descriptive and made sense in the logical context of legal judgments. Controlling the prompts minimized the incidence of hallucination by restricting the output.LexiAI, when compared to baseline methods, scored well for its interpretability and grounding. It achieved a good balance between semantic retrieval and controlled generation, fitting appropriately for a professional-level legal research task.

V. CONCLUSION

LexiAI has proposed its approach of Retrieval-Augmented Generation, which is applicable to the structured summarization of legal cases and document-based legal question answering. Such an approach has further scope in enabling accurate and transparent factual information obtained out of legal research. The proposed approach has immense scope in enabling more efficient and accurate legal information analysis.

In addition, LexiAI addresses the major limitations of traditional keyword-based legal research systems by incorporating semantic understanding and contextual retrieval mechanisms. By integrating transformer-based embeddings with a retrieval-first framework, the system ensures that responses are grounded in authoritative judicial sources rather than relying solely on parametric memory. This significantly reduces the risk of hallucinated content and enhances the credibility of generated outputs.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” in Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017.
[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” in Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS), 2020.
[3] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, “Legal-BERT: The Muppets Straight Out of Law School,” in Findings of the Association for Computational Linguistics (EMNLP Findings), 2020.
[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), Minneapolis, Minnesota, USA, 2019.
[5] M. J. Bommarito and D. M. Katz, “LexNLP: Natural Language Processing and Information Extraction for Legal and Regulatory Texts,” Journal of Open Source Software, vol. 3, no. 27.
[6] X. Zheng, Z. Guo, Z. Qin, Y. Sun, and X. Liu, “Legal Judgment Prediction via Topological Learning,” in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502067|

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in Proceedings of the International Conference on Learning Representations (ICLR), 2013.
[8] I. Beltagy, K. Lo, and A. Cohan, "Longformer: The Long-Document Transformer," in Proceedings of ACL, 2020.
[9] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," in Proceedings of NeurIPS, 2019.
[10] S. Paul, P. Goyal, and S. Sarkar, "CaseSumm: A Benchmark Dataset for Long Legal Document Summarization," in Proceedings of COLING, 2020.
[11] A. Gupta, A. Shrivastava, and S. Palshikar, "Automatic Summarization of Legal Documents using Transformers," in Proceedings of ICAIL, 2021.
[12] X. Li, Y. Sun, J. Zhang, and X. Zhang, "Retrieve, Read, Rerank: Towards End-to-End Multi-Document Reading Comprehension," in Proceedings of ACL, 2018.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details