ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25dec1333

Enhanced Legal Information Access in Nigeria: A Novel Retrieval Augmented Generation (RAG) Approach

[1]Echeonwu, Emmanuel Chinyere;
[2]Bolou, Dickson Bolou; [3]Omonijo, Oluwaseyi Oluwatola;
[4]Ugbogbo, Mike Johnon; [5]Omejieke, Chinenye Ekene

[1]Department of Computer Science.
Nnamdi Azikiwe University
Awka, Anambra State, Nigeria

[2,3,4]Department of Computer Science.
Nigeria Maritime University
Okerenkoko, Delta State, Nigeria

[5]Department of Computing and Mathematics.
Manchester Metropolitan University
United Kingdom

Publication Date: 2025/12/29

Abstract: This study presents a novel Retrieval Augmented Generation (RAG), a text-based query system, for efficient access to Nigerian legal information. Utilizing the Nigerian Constitution and Criminal Code as its knowledge base, the system employs a pipeline involving semantic segmentation, Sentence Transformer embeddings, and vector database indexing for optimized information retrieval. User queries are refined by a Google Gemini large language model, trained as a Nigerian legal expert, to identify key terms and intent before searching the database for the top ten most relevant document chunks. These chunks, along with the refined query and keywords, are then fed back into Gemini to generate a detailed, referenced answer. The current implementation is evaluated using the precision. Recall, F1Score, perplexity and diversity metrics, and results fall within acceptable benchmarks of mean values (0.65, 0.73, 0.68, 14.42, 0.87) respectively, representing a significant advancement in making complex legal big data accessible.

Keywords: *Retrieval Augmented Generation, Embeddings, Bigdata, Vector Database, Large Language Model.*

How to Cite: Echeonwu, Emmanuel Chinyere; Bolou, Dickson Bolou; Omonijo, Oluwaseyi Oluwatola; Ugbogbo, Mike Johnon; Omejieke, Chinenye Ekene (2025) Enhanced Legal Information Access in Nigeria: A Novel Retrieval Augmented Generation (RAG) Approach. *International Journal of Innovative Science and Research Technology*, 10(12), 1822-1825. https://doi.org/10.38124/ijisrt/25dec1333

I. INTRODUCTION

The pivotal role of the system that accesses legal information plays the maintenance and enforcement of the rule of law, as well as the promotion of equitable justice which is an indisputable fact that has long been recognized across various societies. Without the necessary legal knowledge, individuals and entities may find themselves in situations where they are unable to understand their rights, obligations, and the legal procedures that govern their actions. This can lead to a plethora of issues including; but not limited to, the infringement of rights, the perpetuation of injustice, and a general lack of trust in the legal system (United Nation, 2006). In the context of the vast and continuously developing legal terrain of Nigeria, where the legal framework is a rich collection of diverse historical, cultural, and political influences, the challenge of navigating the intricacies of legal databases becomes particularly pronounced. The sheer complexity of these databases, coupled with the often disorganized and extensive nature of legal documents, tend to

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25dec1333

pose significant hurdles for legal practitioners and the lay public alike, thereby underscoring the criticality of developing innovative mechanisms to facilitate the ease of access to such information.

The complexity of the Nigerian legal system is a multifaceted issue that arises from various factors. Firstly, the legal framework in the country is derived from a combination of English common law, statutes, local laws, and customary laws, which have evolved over time and are subject to continuous adaptation and interpretation (Gwangndi, 2016). This diversity in legal sources contributes to the voluminous and sometimes unwieldy nature of the legal documents that are part of the system. Secondly, the rate at which new laws and regulations are enacted, coupled with the dynamic nature of the information. Therefore, this work introduces the Retrieval Augmented Generation (RAG) approach that leverages the power of Artificial Intelligence to surmount these challenges, thereby enhancing the overall legal information landscape in Nigeria. The proposed solution, which is predicated upon the innovative concept of Retrieval Augmented Generation (RAG), represents a paradigm shift in the way legal information is sought, retrieved, and disseminated within the country. The RAG model is specifically designed to cater to the unique characteristics of the legal domain, where the need for precision, relevance, and comprehensiveness in information retrieval is paramount. By integrating advanced algorithms capable of navigating the intricate corpus of legal texts (Vijit et al., 2021). This approach directly aims to provide users with a more intuitive and user-friendly experience, thereby democratizing access to the legal knowledge that is essential for the effective functioning of any society governed by the rule of law.

II. BACKGROUND

The advent of pre-trained neural language models has demonstrated their remarkable capability to absorb and assimilate a significant depth of knowledge from the data they are exposed to (Petroni et al., 2019). These sophisticated systems manage to encapsulate vast amounts of information within their intricate structures, effectively functioning as parameterized implicit knowledge bases that are not dependent on external memory systems for their learning (Raffel et al., 2019; Roberts et al., 2020). This advancement in the field of artificial intelligence has undeniably sparked enthusiasm and intrigue due to its potential for a multitude of applications. However, it is important to recognize that despite these achievements, such models are not devoid of limitations. One of the primary drawbacks associated with these systems is their inherent inability to effortlessly expand or revise their internal memory banks. This inflexibility can pose challenges in dynamic environments where information is continuously evolving or when the need arises to update the model with new facts or correct errors that have been integrated during the learning process. The RAG systems are designed to enhance the understanding and generation process by incorporating external text documents into the mix (Lewis et al., 2021). These models operate with two main components: (i) a retriever component, denoted by $p_\eta(z|x)$, which is equipped with parameters η and it is responsible for providing a ranked list of top-K relevant passages z when presented with a query x, and (ii) a generator component, governed by parameters θ, which constructs the target sequence y by considering the context of the previously generated tokens y_0, the original input x, and the selected passage z (Lewis et al., 2021).

The process of partitioning text into adjacent sections in accordance with its semantic framework, commonly referred to as text segmentation, has persisted as a significant challenge within the realm of linguistic comprehension (Omri et al., 2018). Prior research in this field has predominantly centered around unsupervised learning approaches, including clustering algorithms and graph traversal techniques, largely because of the scarcity of annotated data available for training purposes. Currently, text segmentation can be presented as a supervised learning problem (Omri et al. 2018), and introduces a substantial novel corpus designed explicitly for the purpose of advancing research in this domain.

- *Sentence Embedding*

Sentence embedding translates sentences into a multidimensional vector space, wherein, semantically analogous sentences are positioned in proximity to one another. This approach facilitates the computationally effective evaluation of sentence similarity by employing metrics such as cosine similarity. Traditionally, sentence embeddings were created using methods like: Averaging word embeddings, such as GloVe embeddings, Training an encoder-decoder model, like Skip-Thought, to predict surrounding sentences and Training a Siamese network on labeled data, such as InferSent, which utilizes the Stanford Natural Language Inference (SNLI) dataset (Reimers and Gurevych, 2009).

Reimers and Gurevych (2009) noted that sentence embedding offers the following benefits: efficiency, enhanced semantic representation and transferability to downstream tasks. While sentence embeddings primarily aim to capture semantic similarity, Reimers and Gurevych (2009) emphasized that BERT embeddings can be overly influenced by lexical similarity, sometimes resulting in inaccurate representations of the true semantic relationship between sentences. BERT-flow helps to mitigate this problem by transforming the embeddings into a more isotropic space, reducing the excessive correlation between lexical and semantic similarity (Li et al., 2020).

- *Vector Database Indexing*

A vector database is optimized to manage and process high-dimensional vectors effectively. These vectors are derived from various data types, including text, images, audio, and video, by applying specialized transformation techniques known as embedding functions. This allows the database to efficiently deal with the complex mathematical representations of data features (Han et al., 2023). Nearest neighbor search, often referred to as similarity search, plays a crucial role in the functionality of vector databases. Han et al. (2023) opined that its primary objective is to identify and retrieve data points that exhibit the highest degree of similarity to a specified query point. This process is especially significant when dealing with high-dimensional data, which

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25dec1333

poses unique challenges that typical database management systems (DBMS) may struggle to address efficiently. The integration of Large Language Models (LLM) and vector databases is a promising area with the potential to revolutionize how we interact with information and create more intelligent systems.

III. METHODS

The approach used are broken down into different areas such as segmentation, embeddings, vector indexing, query refinement, retrieval and generation. Legal texts, such as the Nigerian Constitution and Criminal Code, are known for their extensive length and intricate nature. To facilitate better comprehension and retrieval of information, semantic segmentation techniques were applied. This method involves partitioning the documents into meaningful sections that are coherent from a semantic standpoint. It offers a more nuanced approach compared to the conventional methods that are limited to analyzing sentences or paragraphs, which may sometimes fail to provide the necessary contextual depth. The user queries, often expressed in natural language, were refined to improve the accuracy of the retrieval process. A Google Gemini LLM, trained on a large corpus of Nigerian legal documents, acts as a legal expert to give semantic understanding. This LLM analyzes user queries to identify key terms, disambiguate ambiguous language, and understand the underlying intent behind the query. The refined query is used to search the vector database, retrieving the top ten most relevant document chunks. These chunks, along with the refined query and extracted keywords are then fed into the Gemini LLM. The LLM synthesizes this information to generate a detailed and referenced answer, ensuring both accuracy and transparency.

A. *The RAG System Development Cycle*

➢ *Segmentation:*
- Break down the input data into smaller, manageable pieces.
- This could involve dividing text into sentences or paragraphs.

➢ *Embedding:*
- Convert the segments into a numerical format that captures their semantic meaning.

➢ *Vector Indexing:*
- Organize the embedded vectors into a structured format that allows for efficient searching and retrieval.

➢ *Query Refinement:*
- Improve the user's query to better match the indexed data.
- Expanding the query with synonyms, using query expansion techniques, or re-weighting terms based on their importance.

➢ *Retrieval:*
- Search the indexed vectors to find the most relevant segments based on the refined query.
- Use Retrieval method: cosine similarity

➢ *Generation:*
- Use the retrieved information to generate a response using an LLM.
- In the case of text, this could involve summarizing the retrieved segments or creating a coherent narrative.

IV. RESULTS AND DISCUSSION

BERTScore (Precision, Recall, and F1Score) uses BERT embeddings to determine the semantic similarity between the generated output and reference text. Range: 0–1. Higher scores indicate greater semantic similarity. Perplexity indicates how effectively a language model anticipates the text. Range from 1 to ∞. Lower levels suggest improved fluency and coherence. Diversity assesses the originality of bigrams in generated output. Range: 0–1. Higher numbers correspond to more diversified and varied output.
The provided metrics in Table 1. offer valuable insights into the performance of the language model used across five questions. The BERTScore metrics, including precision, recall, and F1 score, indicate a strong semantic similarity between the generated output and the reference text, with mean scores of 0.65, 0.73, and 0.68, respectively. These scores suggest that the model is capable of producing output that is contextually relevant and coherent. The F1 score, which balances precision and recall, consistently hovers around 0.68, indicating a stable performance across questions.

The perplexity scores, which measure the system model's ability to anticipate the text, show a moderate level of fluency and coherence, with a mean score of 14.42. While lower perplexity scores are generally desirable, the scores in this dataset are not excessively high, suggesting that the model is able to generate text that is reasonably coherent and easy to follow. However, there is some variation in perplexity scores across questions, with Question 4 exhibiting the highest perplexity score of 17.74, indicating a slightly lower level of fluency and coherence in this specific context.

The diversity metric, which assesses the originality of bigrams in the generated output, reveals a strong performance, with a mean score of 0.87. This suggests that the model is able to produce output that is not only coherent but also varied and diverse. The diversity scores are consistently high across questions, indicating that the model is able to adapt to different contexts and generate novel responses. Overall, the analysis suggests that the language model is performing well in terms of semantic similarity, fluency, and diversity, but there is room for improvement in certain areas, such as reducing perplexity scores to achieve even greater coherence and fluency. While the system shows some promise in terms of diversity, its fluency, coherence, and semantic precision need significant improvement. Focusing on reducing perplexity and improving the precision of the generated text should be the priorities for future development. While topic intersection, high enthropy and textual version collusion could result in high perplexity; hallucination, confabulation, and textual mimicry can also affect the precision of the model's output responses. By analyzing the individual inputs where perplexity is highest might reveal specific weaknesses in the model's handling of certain types of queries.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25dec1333

V. CONCLUSION

This work demonstrates the effectiveness of a RAG-based approach for legal information retrieval within the complex Nigerian legal landscape. It offers a more accessible alternative to conventional methods. While the handling of large and highly structured legal documents remains non-trivial, the system produces encouraging results. Ongoing efforts are directed toward refining the approach and improving overall system performance and robustness.

The RAG system represents a substantial stride towards making Nigerian legal information more accessible to a wider audience, thereby empowering legal practitioners and the public alike with a more efficient tool for navigating the legal system's complexities. As we continue to develop and refine this technology, we anticipate substantial contributions to the enhancement of access to justice and the promotion of a more transparent and equitable legal environment across the nation.

Table 1: Summary Evaluation Metrics Used.

Metrics	BERT Precision	BERT Recall	BERT F1Score	Perplexity	Diversity
Question 1	0.65	0.71	0.68	10.1	0.84
Question 2	0.65	0.72	0.68	15.1	0.89
Question 3	0.66	0.76	0.7	12.35	0.87
Question 4	0.64	0.73	0.68	17.74	0.92
Question 5	0.65	0.76	0.7	16.84	0.86
Mean	**0.65**	**0.73**	**0.68**	**14.42**	**0.87**

REFERENCES

[1]. Gwangndi, M., I. (2016). The Socio-Legal Context of the Nigerian Legal System and the Shariah Controversy: An Analysis of Its Impact on Some Aspects of Nigerian Women'S Rights. Journal of Law, Policy and Globalization. 45: 2224-3240.

[2]. Han, Y., Liu, C., and Wang, P., (2023). A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge. https://arxiv.org/pdf/2310.11703.

[3]. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.*ArXiv, abs/2005.11401*.

[4]. Li, B., Zhou, H., He, J., Wang, M., Yang, Y., and Li, L., (2020). On the Sentence Embeddings from Pre-trained Language Models. Art. no. arXiv:2011.05864, 2020. doi:10.48550/arXiv.2011.05864.

[5]. Omri K., Adir C., Noam, M., Rotman, M., and Berant, J. (2018).Text Segmentation as a Supervised Learning Task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers).* pages 469–473, New Orleans, Louisiana. Association for Computational Linguistics.

[6]. Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y. and Miller, A. (2019). Language models as knowledge bases? In Proceedings of the *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* pages 2463–2473. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://www.aclweb.org/anthology/D19-1250.

[7]. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv e-prints. URL https://arxiv.org/abs/1910.10683.

[8]. Reimers, N. and Gurevych, I., (2009). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Art. no. arXiv:1908.10084, doi:10.48550/arXiv.1908.10084.

[9]. Roberts, A., Raffel, C., and Shazeer, N. (2020). How much knowledge can you pack into the parameters of a language model? arXiv e-prints. URL https://arxiv.org/abs/ 2002.08910.

[10]. United Nation. (2006). Social Justice in an open world. Publication by the Department of Economic and Social Affairs. International Forum for Social Development. https://www.un.org/esa/socdev/documents/ifsd/SocialJustice.pdf

[11]. Vijit M., Rishabh, S., Kumar G., Shubham K, M.,, Angshuman H., Arnab B., Ashutosh, M. (2021). Semantic Segmentation of Legal Documents via Rhetorical Roles. Art. no. arXiv:2112.01836, doi:10.48550/arXiv.2112.01836.