Innovative Informatics and Artificial Intelligence Research (IIAIR)
Vol. 1, Issue 1, pp. 35-41, 2025
Received 19 Jan 2025; accepted 23 March 2025; published 30 Apr 2025
http://doi.org/10.35718/iiair.v1i1.1306

Retrieval-Augmented Generation for Indonesian Criminal Law Information Using the LLaMA Model

Ariel Zakly Pratama[1], Arwin Marinta[2], Bagus Triyudanto[3] ,Muhammad Saman[4], Tirana Noor Fatyanosa[5]

[1, 2, 3, 4]Informatics, Kalimantan Institute of Technology, Balikpapan, Kalimantan Timur, Indonesia.
[5]Universitas Brawijaya, Malang, Jawa Timur, Indonesia.

Corresponding author: Tirana Noor Fatyanosa (fatyanosa@ub.ac.id)

To cite this article: A. Z. Pratama, A. Marinta, B. Triyudanto, M. Saman, T. N. Fatyanosa, “Retrieval-Augmented Generation for Indonesian Criminal Law Information Using the LLaMA Model,” *Innovative Informatics and Artificial Intelligence Research*, vol. 1, issue 1, 2025. [Online]. Available: https://doi.org/10.35718/iiair.v1i1.1306

Tirana Noor Fatyanosa serves as an Editor of IIAIR but was not involved in the peer-review process of this article

Abstract

Law is a set of rules that regulate boundaries of behavior within society, is coercive in nature, and imposes sanctions on violators. One of its branches is criminal law, which focuses on violations of norms with the threat of sanctions. However, the legal system often provides excessive discretion to the judiciary, making legal outcomes difficult to predict and potentially disadvantageous, especially for individuals who lack legal understanding or cannot afford legal representation. This disparity poses a significant challenge to legal systems in many countries. In this study, we demonstrate that artificial intelligence (AI) based on Retrieval-Augmented Generation (RAG) can serve as an innovative solution to support a more equitable enforcement of the law. This technology integrates information retrieval and data-driven text generation to help the public understand their rights, access valid legal information, and obtain relevant legal guidance. Based on the implementation and testing of a criminal law chatbot using the LLaMA language model, questions, generated answers, and the relevance of the chatbot's responses were evaluated. Out of ten tested questions, eight received relevant responses, while two were deemed irrelevant. Additionally, the chatbot's capability to retrieve legal documents based on user-provided prompts was assessed. Of ten input prompts, the chatbot successfully identified eight relevant documents, achieving a hit rate of 80%. Thus, the application of RAG in this legal chatbot can provide an innovative solution to support law enforcement. In the future, the use of AI in legal systems has the potential to reduce information disparities, enhance transparency in legal processes, and create a more efficient and accessible legal system for all.

Keywords: artificial intelligence; deep learning; chatbot; RAG; law; LLM; LLaMA

1. Introduction

Law is a set of rules that applies within society, functioning to regulate behavioral boundaries. It is coercive in nature and imposes sanctions on violators. Law governs relationships within society by providing rights and obligations. Given that interactions between individuals can often lead to conflicts, laws are created as a means of mitigating these conflicts as effectively as possible [1].

Law encompasses many branches, one of which is criminal law. The term "criminal" originates from the concept of suffering or punishment, which implies that punishment is imposed on individuals who violate the law. Two critical elements in criminal law are norms and sanctions: norms define prohibitions or commands, while sanctions are the penalties imposed on those who violate these norms [2].

Society as a whole relies on laws to regulate life. However, the legal system often provides excessive discretion to the judiciary, making outcomes unpredictable. This unpredictability frequently leads to individuals experiencing losses, both financially and in terms of personal freedom [3].

To address these challenges, every individual in society must gain a proper understanding of the law to defend themselves effectively when facing legal issues. Many individuals, however, lack legal knowledge and must rely on lawyers to represent them in court. This situation differs significantly for those who are less fortunate; individuals who cannot afford legal counsel and lack legal understanding may be punished for offenses they did not commit. This disparity in legal access is a common issue in many countries worldwide.

With technological advancements, artificial intelligence (AI) has emerged as a breakthrough solution to

This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY)

address legal disparities. AI, which mimics human intelligence in data processing, offers significant potential to support transparent and efficient law enforcement.

Therefore, to support law enforcement and help society access valid legal information, this study will develop a chatbot using the RAG approach. This will enable the chatbot to retrieve external legal knowledge from legislative documents or any official legal sources in the country. By leveraging this capability, chatbot users can access legal information and study it independently, allowing them to conduct legal defense when needed.

2. Related Works

Artificial Intelligence (AI) is increasingly applied in the legal field, with significant advancements in chatbot systems and natural language processing (NLP), especially with the advancement of transformer [4]. In 2021, deep learning was utilized in the development of chatbots for medical purposes, highlighting the flexibility of chatbot technology across various domains [5]. In the same year, an AI-based legal consultation chatbot system was developed, showcasing AI's potential to provide accessible legal assistance and advice [6]. By 2023, research expanded this concept to more complex legal tasks, such as multi-intent classification and slot filling for Supreme Court decisions, demonstrating the effectiveness of transformer-based models in legal text analysis [4]. Subsequently, the application of language models was further developed using transformer architectures for AI in law, offering insights into structuring these models to address domain-specific challenges [7].

These studies collectively illustrate the growing intersection between AI, NLP, and the legal field, while providing a foundation for developing chatbot systems tailored to legal applications, particularly in criminal law.

3. Methods

This study employs a quantitative approach aimed at measuring the effectiveness of legal chatbots in providing solutions to legal issues faced by the public. Data will be collected through structured surveys, and statistical analysis will be conducted to test the proposed hypotheses.

3.1 Dataset

Dataset is an organized collection of data designed for various purposes such as analysis and research. Datasets can take different forms, including tabular data, images, text, and audio. For tabular data, the structure consists of columns representing attributes and rows representing records. The dataset used in this experiment was obtained from peraturan.go.id, which includes the Law of the Republic of Indonesia Number 1 of 2021 concerning the Criminal Code (KUHP) as the primary source of information.

3.2 PyMuPDF

PyMuPDF is a document manipulation library available in Python. In this study, PyMuPDF was used to extract text from the Law of the Republic of Indonesia Number 1 of 2021 concerning the Criminal Code (KUHP).

3.3 LlaMA

LLaMA is an example of a Generative Pretrained Transformer (GPT) model developed based on the original Transformer architecture [8]. Transformer-based models, such as LLaMA, are recognized as high-performing models in the field of Natural Language Processing (NLP). In these models, the output is generated through vocabulary classification at each generative step, based on a sequence of latent representations obtained from the processes of tokenization and embedding [9].

LLaMA adopts pre-normalization methods similar to GPT-3, utilizing the RMS Norm normalization function on the input of each transformer sub-layer. This approach enhances training stability by adjusting invariance properties and automatically adapting to learning rates. Additionally, it replaces the conventional ReLU activation function with SwiGLU, which provides more optimal training performance. LLaMA also incorporates rotary positional embeddings (RoPE) in each layer, contributing to improved model performance [4].

3.4 Word Embedding

Word embedding is a technique that transforms text into numerical representations, also known as word representation techniques. Word embedding plays a crucial role in text mining because machine learning techniques cannot operate directly on text-based data. Technically, word embedding converts each word into a numerical representation using a specific vocabulary. This technique can be trained on large text corpora by applying neural networks.

3.5 Retrieval Augmented Generation

LLMs have demonstrated strong capabilities in understanding instructions and generating text. However, many studies have found that LLMs often experience hallucinations due to outdated information or inaccurate knowledge. Large-scale, uncontrolled data collection, the low proportion of high-quality data, imperfections in data allocation within the input space, as well as various other factors, can exacerbate this issue [10].

Retrieval Augmented Generation (RAG) is a method that integrates information retrieval processes, allowing the generation process to retrieve relevant objects from available data sources in order to improve accuracy and precision, thereby reducing hallucinations.

The foundations of RAG are classified based on how the retriever strengthens the generator, forming a fundamental abstraction of augmentation methodologies for various types of retrievers and generators. This unified perspective encompasses all RAG scenarios and explains key advancements and technologies that support future developments. Additionally, summarization methods are introduced as enhancements to RAG, facilitating more effective engineering and implementation of RAG systems.

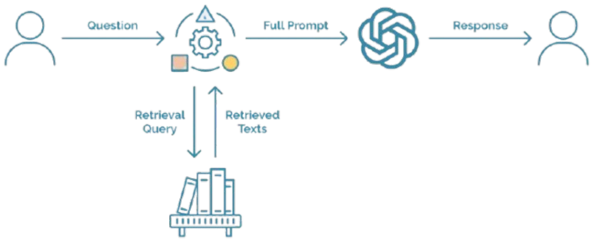

Figure 1: A workflow of RAG [11]

As shown in Figure 1, RAG works by receiving input from the user. The input is then used as a query to search for relevant documents, after which the appropriate text is retrieved and processed by a large language model (LLM) to generate an interactive response.

3.6 ChromaDB

ChromaDB is an open-source database designed to efficiently store and retrieve vector representations. In the case of RAG-based chatbots, ChromaDB can be used to store documents that serve as external knowledge for the LLM [12]. The main advantage of ChromaDB lies in its ability to store the database directly on the user's device, thereby offering greater flexibility and control over data management and access [13].

ChromaDB creates a local database, referred to as a *collection*. Within a collection, documents can be stored along with their descriptions and metadata to support the management of the vector database. ChromaDB can be integrated with a local embedding model, allowing automatic generation of embeddings when documents are added to the collection, with the embeddings subsequently stored within the same collection. Furthermore, ChromaDB is capable of handling vector similarity searches, a feature that enables retrieval of documents relevant to the user’s input prior to generating a response [14].

3.7 Evaluation Metric

Evaluation metrics offer advantages such as low cost and ease of use for comparing different systems, but their quality has yet to match human evaluation. To address this limitation, research continues to develop automatic evaluation metrics that can mimic or even replace manual evaluation methods. Automatic evaluation metrics are performed by comparing machine translation results with reference translations that are considered correct [15].

Hit Rate is a binary metric that determines whether there is a relevant item in the recommendation list for a user. If a relevant item is found in the recommendation list, it is considered a hit [16]. This metric is used to measure how often a recommendation system successfully suggests items that match the user's needs. The higher the hit rate, the more relevant the recommendation list is. The hit rate formula is shown in Equation 1.

$$\text{Hit Rate} = \frac{\text{Number of Relevant Items Retrieved}}{TotalNumberofRelevantItems} \tag{1}$$

4. Experimental Settings

This study aims to develop and evaluate the performance of a criminal law assistant chatbot model using the Retrieval-Augmented Generation (RAG) approach with the pretrained LLaMA model. The model's performance is measured based on two main metrics: the relevance of the answers and the accuracy of the information provided in relation to criminal law context. The model is designed to provide answers to questions related to Indonesian criminal law by integrating two main stages: retrieving relevant legal documents and generating answers based on the retrieved context.

The dataset used in this experiment is obtained from peraturan.go.id, which includes the Republic of Indonesia Law Number 1 of 2021 on the Criminal Code (KUHP) as the primary source of information.

This data undergoes a chunking process, where the document is divided by articles. This is done by searching for text starting with the phrase “Pasal ”, and the chunk ends when another “Pasal ” is encountered. Additionally, text attributes such as ‘\n’ and ‘\x01’ are replaced with spaces (‘ ‘) to clean up the text.

After chunking, the chunks are stored in ChromaDB for embedding storage. Each chunk is also assigned metadata before being saved to facilitate document retrieval. The metadata used includes a “label” with the value “Kitab Undang-Undang Hukum Pidana Tahun 2021,” and “pasal_number” derived from the chunk. This process aims to improve the model’s accuracy in understanding the context of prompts for document retrieval.

5. Results and Discussions

In this section, we will discuss the implementation and testing results of the criminal law chatbot developed using the LLaMA language model. The purpose of developing this chatbot is to provide quick and accurate legal information in the context of criminal law, leveraging natural language processing technology to answer questions related to regulations, legal procedures, and criminal cases.

The testing will be conducted by providing 10 questions to the chatbot model. After that, the relevance of the chatbot's responses will be evaluated based on the original source of the answers. The results of the testing on this criminal law chatbot can be seen in the table below.

Tabel 1: The reuslts of the text generation derived from LLaMa

No	Questions	Answers	Is Relevant?
1	Bagaimana cara seseorang bisa dilepaskan bersyarat?	Berdasarkan informasi yang diberikan, ada beberapa cara seseorang bisa dilepaskan bersyarat. Berdasarkan Pasal 77, jika tertuduh meninggal dunia, maka seseorang bisa dilepaskan hapus dari tuntutan pidana. Namun, terdapat satu syarat yaitu tertuduh harus meninggal dunia.	✓
2	Apa hukuman pencurian?	"Pasal 448 Seorang penumpang kapal Indonesia yang	✓

		merampas kekuasaan atas kapal secara melawan hukum, diancam dengan pidana penjara paling lama tujuh tahun." Jadi, hukuman pencurian kapal adalah penjara selama tujuh tahun.	
3	Apa hukuman pada orang yang memperdagangkan manusia?	Berdasarkan informasi 1, pasal 297 Perdagangan wanita dan perdagangan anak laki-laki yang belum dewasa diancam dengan pidana penjara paling lama enam tahun.	✓
4	Orang yang melukai sebagai bentuk pembelaan diri, apakah dipidana?	Berdasarkan informasi yang diberikan, orang yang melakukan pelangaran dan kemudian menjadi korban sendiri (melukai sebagai bentuk pembelaan diri) tidak akan dipidana. Hal ini disebut dalam Pasal 60 yang menyatakan bahwa membantu melakukan pelangaran tidak dipidana. Jadi, jawaban dari pertanyaan tersebut adalah: Tidak, orang yang melukai sendiri sebagai bentuk pembelaan diri tidak akan dipidana.	✓
5	Apakah percobaan pelanggaran dapat dipidana?	Berdasarkan Informasi 1 yang menyatakan "Pasal 60 Membantu melakukan pelanggaran tidak dipidana", maka jawaban dari pertanyaan tersebut adalah: Ya, karena pasal ini menyatakan bahwa membantu melakukan pelanggaran tidak dipidana.	✓
6	Apa hukuman bagi orang yang menyebarkan surat rahasia?	Berdasarkan Pasal 554, hukuman bagi orang yang tanpa izin penguasa yang berwenang menahan surat-surat jabatan adalah pidana kurungan paling lama dua bulan atau pidana denda paling banyak empat ribu lima ratus rupiah.	✓
7	Apa hukuman bagi pemberontak?	Berdasarkan informasi yang diberikan, tidak ada informasi langsung tentang hukuman bagi pemberontak. Namun, dapat disimpulkan bahwa Pasal 83 dan Pasal 89 berhubungan dengan pidana hapus dan kekerasan. Pasal 88 menyebutkan tentang permufakatan jahat, namun tidak secara eksplis dihubungkan dengan hukuman bagi pemberontak. Namun, karena Pasal 89 tidak diterapkan, maka tidak ada informasi tentang hukuman bagi pemberontak.	✗

8	Apa hukuman bagi orang yang mencoba merampas kemerdekaan?	Berdasarkan informasi yang diberikan, tidak ada pasal yang secara langsung menyebutkan tentang "mencoba merampas kemerdekaan". Namun, dapat dipahami bahwa "kemerdekaan" dalam konteks ini mungkin merujuk pada hak asasi manusia atau kebebasan individu. Dalam hal ini, tidak ada informasi yang menyebutkan tentang hukuman bagi orang yang mencoba merampas kemerdekaan. Namun, dapat dipahami bahwa jika kita mempertimbangkan pasal-pasal yang terkait dengan hak asasi manusia dan kebebasan individu, maka hukuman yang diberikan oleh Pasal 23 dapat digunakan sebagai acuan. Pasal 23 menyebutkan bahwa orang yang dijatuhi pidana kurungan dapat "meringankan nasibnya menurut aturan-aturan yang akan ditetapkan dengan undang-undang". Dalam konteks ini, hukuman yang diberikan mungkin berupa penanganan yang lebih lembut atau penyesuaian dengan kebutuhan individu. Namun, perlu diingat bahwa Pasal 23 tidak menyebutkan secara eksplisit tentang "mencoba merampas kemerdekaan" sebagai tindakan yang dapat dilakukan. Oleh karena itu, hukuman yang diberikan oleh Pasal 23 mungkin tidak berlaku dalam konteks ini. Dalam hal ini, tidak ada jawaban yang pasti dan spesifik tentang hukuman bagi orang yang mencoba merampas kemerdekaan.	✗
9	Apa hukuman menyerang kehormatan orang lain?	Berdasarkan informasi yang disediakan, hukuman untuk menyerang kehormatan orang lain dapat ditentukan dengan memandang Pasal 281. Menurut Pasal 281 ayat (1) dan (2), barang siapa dengan sengaja dan terbuka melanggar kesusilaan akan dihukum dengan pidana penjara paling lama dua tahun delapan bulan atau pidana denda paling banyak empat ribu lima ratus rupiah. Sementara itu, pasal (2) mengatakan bahwa barang siapa dengan sengaja dan di depan orang lain yang ada di	✓

This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY)

		situ bertentangan dengan kehendaknya, melanggar kesusilaan pun akan dihukum dengan pidana penjara paling lama dua tahun delapan bulan atau pidana denda paling banyak empat ribu lima ratus rupiah.	
10	Apa hukuman bagi pejabat yang menyalahgunakan kekuasaan?	Berdasarkan Informasi 4, pasal 421 menyatakan bahwa seorang pejabat yang menyalahgunakan kekuasaannya diancam dengan pidana penjara paling lama dua tahun delapan bulan.	✓

Table 1 shows the testing results of the developed chatbot. The table includes the questions, the answers generated, and the relevance of the answers provided by the chatbot. Questions one through ten are taken from the dataset used as the source of the test data. It is observed that eight out of ten (80%) responses provided by the chatbot are relevant.

Tabel 2: Retrieval performance result

Evaluation Metric	**Input**	**Score**
Hit Rate	Questions in Tabel 1	80%

Table 2 presents the results obtained, showing that the LLaMA-based chatbot can effectively search for documents based on the prompts provided by the user. Out of the 10 prompts tested, the chatbot successfully found 8 out of 10 documents related to the given prompts, achieving a success rate of 80%.

6. Conclusions

The integration of Artificial Intelligence (AI) technology in law enforcement, particularly through the Retrieval Augmented Generation (RAG) approach, offers significant potential for creating a more transparent and efficient legal system. With technological advancements, AI is now capable of addressing major challenges related to accessibility and understanding of the law, which are often barriers for the public when dealing with legal issues. By utilizing RAG, which combines data-driven information retrieval with the ability to generate relevant text, this system can provide accurate answers based on valid documents, a crucial aspect in the context of criminal law.

In this research, the application of RAG technology to develop an AI-based chatbot that can provide more precise and relevant legal answers has proven effective. The system, which accesses and processes legal information from various documents such as the Indonesian Criminal Code (Kitab Undang-Undang Hukum Pidana), is able to filter information according to the user's query. By using metadata that includes labels and accurate legal sources, the system not only enhances the relevance of the answers but also ensures that the information provided comes from legitimate and up-to-date texts. This process demonstrates how technologies such as ChromaDB and language models like LLaMA 3.2 can be integrated to improve the quality of interaction between the public and the legal system.

Thus, the application of RAG in this legal chatbot can serve as an innovative solution to support law enforcement. It can help the public easily access legal information, understand their rights, and receive accurate guidance on applicable legal procedures.

Furthermore, this model can serve as a reference for the implementation of RAG-based chatbot models in other fields. As long as there are documents that can be used as knowledge sources, RAG-based chatbots can be implemented. Through this research, it is hoped that it will contribute to the study and development of automated services such as chatbots in any field.

Declaration of Conflicting Interests

The authors declare that there are no competing interests that could have influenced the work of our study.

References

[1] A. E. N. Lubis and F. D. Fahmi, "PENGENALAN DAN DEFINISI HUKUM SECARA UMUM (LITERATURE REVIEW ETIKA)", JIMT, vol. 2, no. 6, pp. 768–789, Jul. 2021.

[2] A. Efritadewi, "Modul Hukum Pidana. Tanjungpinang–Kepulauan Riau,", Universitas Maritim Raja ali Haji, 2020

[3] O. W. Holmes Jr, "The common law," Routledge, 2020.

[4] A. Munthuli et al., "Transformers for Multi-Intent Classification and Slot Filling of Supreme Court Decisions Related to Sexual Violence Law," in IEEE Access, vol. 11, pp. 76448-76467, 2023, doi: 10.1109/ACCESS.2023.3296261.

[5] V. Tode et al., "A chatbot for medical purpose using deep learning," International Journal of Engineering Research & Technology, 2021

[6] ParkKoo-Rack, 'Development of Artificial Intelligence-based Legal Counseling Chatbot System', Journal of the Korea Society of Computer and

Information, vol. 26, no. 3, pp. 29–34, 03 2021.

[7] C. M. Greco and A. Tagarelli, “Bringing order into the realm of Transformer-based language models for artificial intelligence and law,” Artificial Intelligence and Law, vol. 32, no. 4. Springer Science and Business Media LLC, pp. 863–1010, Nov. 20, 2023. doi: 10.1007/s10506-023-09374-7.

[8] K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, “Llama 2: Early Adopters’ Utilization of Meta’s New Open-Source Pretrained Model.” MDPI AG, Aug. 01, 2023. doi: 10.20944/preprints202307.2142.v1.

[9] P. Zhao et al., “Retrieval-Augmented Generation for AI-Generated Content: A Survey,” 2024, arXiv. doi: 10.48550/ARXIV.2402.19473.

[10] S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling, “Corrective Retrieval Augmented Generation,” 2024, arXiv. doi: 10.48550/ARXIV.2401.15884.

[11] “What is RAG (Retrieval Augmented Generation)?,” Trantorinc, May 07, 2024. https://www.trantorinc.com/blog/what-is-rag-retrieval-augmented-generation

[12] T. Taipalus, “Vector database management systems: Fundamental concepts, use-cases, and current challenges,” Cognitive Systems Research, vol. 85. Elsevier BV, p. 101216, Jun. 2024. doi: 10.1016/j.cogsys.2024.101216.

[13] S. Alshammari, L. Basalelah, W. A. Rukbah, A. Alsuhibani, and D. S. Wijesinghe, “PyZoBot: A Platform for Conversational Information Extraction and Synthesis from Curated Zotero Reference Libraries through Advanced Retrieval-Augmented Generation,” 2024, arXiv. doi: 10.48550/ARXIV.2405.07963.

[14] J. Knapp and K. Moberg, “From Data to Descriptions: Efficient Data Retrieval in Autonomous Vehicle Development using Generative AI,” 2024

[15] S. Lee et al., “A Survey on Evaluation Metrics for Machine Translation,” Mathematics, vol. 11, no. 4. MDPI AG, p. 1006, Feb. 16, 2023. doi: 10.3390/math11041006.

[16] H. H. Arfisko and A. T. Wibowo, “Sistem Rekomendasi Film Menggunakan Metode Hybrid Collaborative Filtering Dan Content-based Filtering,” eProceedings of Engineering, vol. 9, no. 3, 2022.

This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY)