Enhanced Large Language Models-based Legal Query Responses through Retrieval Augmented Generation

Ruiteng Li

[1] International Energy College, Jinan University, Zhuhai, 519000, China
email: kurosakana@stu2021.jnu.edu.cn

Abstract. This study aims to solve the challenge of hallucination in Large Language Models (LLMs) through the implementation of Retrieval Augmented Generation (RAG). Especially the case of answering legal questions. Using Canada's Criminal Code as dataset, the research first converts the code to the text file format. Then segmenting the text into vectors for a vector database, and with the help of an Embedding Model for vector transformation. Finally, the LLM will play the role of search engine and provide answer. As a result, the evaluation using the Automated Evaluation of Retrieval Augmented Generation (RAGAS) method demonstrated high faithfulness of 0.9444 and answer relevancy of 0.7488, indicating the model's factual accuracy and relevance to the posed questions. Despite the complexity of the legal questions and the initial challenge of understanding indirect queries, the RAG model successfully provided accurate and relevant answers. However, context precision and context recall were relatively lower at 0.4 and 0.45, suggesting potential for improvement in the model's ability and the evaluation model's possible deficiency. The study successfully demonstrates the potential of RAG to mitigate hallucination issues in LLMs, particularly in the legal domain. The high faithfulness and relevancy scores affirm the model's efficacy in providing accurate legal information, marking a significant advancement in the application of LLMs for legal queries. Future research could focus on enhancing context precision and recall, or create more precise database to evaluate with better evaluation methods.

Keywords: Retrieved Augmented Generation (RAG), AI Hallucination, Artificial Intelligence.

1 Introduction

Nowadays, people live in an information explosion era in which people can access all sorts of data to help them improve their lives, especially in the context of rapid development of machine learning models and their applications in many fields such as industries [1-4]. One of the most important pieces of information is legal questions. However, responses found on the Internet vary in quality and are often unreliable. What's more, considering the vast number of laws, even some legal professionals struggle to recall them all. Fortunately, Large Language Models (LLMs) have

Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_27

achieved great breakthrough recently [5]. Therefore, people can ask LLMs to know the answers to their legal problems with natural language. However, LLMs are not perfect, in fact they often give a certain answer while it's fake or wrong. This situation known as hallucination [6] is very misleading and harmful for the users. Yet, this conundrum can be fixed by a method called Retrieval Augmented Generation (RAG) [7]. RAG fixes document representations but allows context representations to update during training to better fit the retriever for the task. Using RAG, hallucination will be mitigated or even eliminated, so the users can get convincing answers.

Not until ChatGPT showed up did most people realize the importance of LLMs. Hence previous research is not abundant though there are some related studies that have been published. For instance, Antoine et al. proposed a dataset, the Long-form Legal Question Answering (LLeQA) dataset, to train LLMs to answer legal questions [8]. But it does not mention providing a specific LLM for answering legal questions. Conversely, this research includes training and applying for a LLM to answer legal questions directly. Another article by Chouhan et al. discusses the LexDrafter framework, which aids in drafting Definitions articles for legislative documents by utilizing RAG and existing term definitions. It aims to standardize legal definitions across various regulations, minimize human errors, and streamline the drafting process [9]. Although it’s more relevant to LLM using RAG to answer legal questions, it’s still not the case. Since it focusses more on term definitions rather than specific questions. While this work aims to deal with actual cases encountered by users. In conclusion, there is no previous work that actually trained a LLM to provide legal questions with reference.

This work was initiated to address the challenge of hallucination in LLMs through the implementation of RAG. To complete this study, Canada criminal code was chosen to be the dataset. The criminal code text originates from one of the Canada’s government websites, namely Justice Law Website. It’s Consolidated Acts part contains this document in HTML XML and PDF format. The criminal code in HTML format was downloaded and then converted into TXT format. Then Embedding Model was used to transform words into vector. Based on Milvus vector database, LLM was trained with vectors generated by Embedding Model. Finally, LLM will interact with people using natural language. It will precisely understand the query’s meaning and use it to search the vector database. Achieving the aim to help normal people and legal professionals to quickly get answers from numerous codes.

2 Method

2.1 Dataset Preparation

The source of the dataset is Canada’s Justice Law Website Criminal Code (R.S.C., 1985, c. C-46) from which its HTML format is downloaded [10]. Then it was converted from HTML format to TXT format after some irrelevant elements was removed since they are interruption to the LLM and meaningless to human. Although the PDF is also very clear, it included French text, so it wasn’t chosen because deleting French text from it will need more time.

The dataset was loaded then was used to segment the content of the text into multiple sentence windows, creating nodes for each window. These nodes include the original text content and related metadata in order to prepare for creating vector database. Embedding Model was used to transform text into numerical vector nodes. These vectors captured the semantic information. Hence, the vectors can be searched by their similarity. Next, the nodes are added into vector storage and construct index, which allows quick retrieval and search for the most related node [11]. The way to retrieve data from the vector database is like search engine searching from the Internet. While the search engine role is played by LLMs, which allows users to get their answers in natural language rather than keywords. The LLM will understand the meaning of the query and search it through vector database.

2.2 RAG

On a related note, users can use natural language to retrieve data thanks to Natural Language Processing (NLP) and LLM. NLP refers to the field of study that focuses on the interaction between computers and human language, how human and computers communicate effectively using natural language. Recent advancements in NLP have led to the development of powerful language models such as the GPT series, including LLM such as ChatGPT. These models are pre-trained on vast amounts of text data and have demonstrated exceptional performance in various NLP tasks such as language translation, text summarization, and question-answering [12]. That's why they are used to retrieve data and answer the user's questions. RAG is the key to let LLMs give precise answer avoiding hallucination, a research paradigm that enhances the capabilities of LLMs by incorporating external knowledge sources.

RAG works by sourcing and integrating information from external databases to provide comprehensive prompts for LLMs to generate well-informed answers. As the name indicated, RAG's framework shown in Fig. 1 comprises three main stages: Indexing, Retrieval, and Generation. 1) Indexing: Raw data is cleaned, extracted, and converted into plain text format. The text is segmented, encoded into vectors, and stored in a vector database for efficient similarity searches during retrieval. 2) Retrieval: User queries are transformed into vector representations, and similarity scores with indexed chunks are computed. Top K chunks with the highest similarity are retrieved for the generation stage. 3) Generation: Queries and retrieved documents are combined to prompt a large language model for response generation. The model can utilize inherent knowledge or restrict responses to provided documents, adapting to task-specific criteria [13].

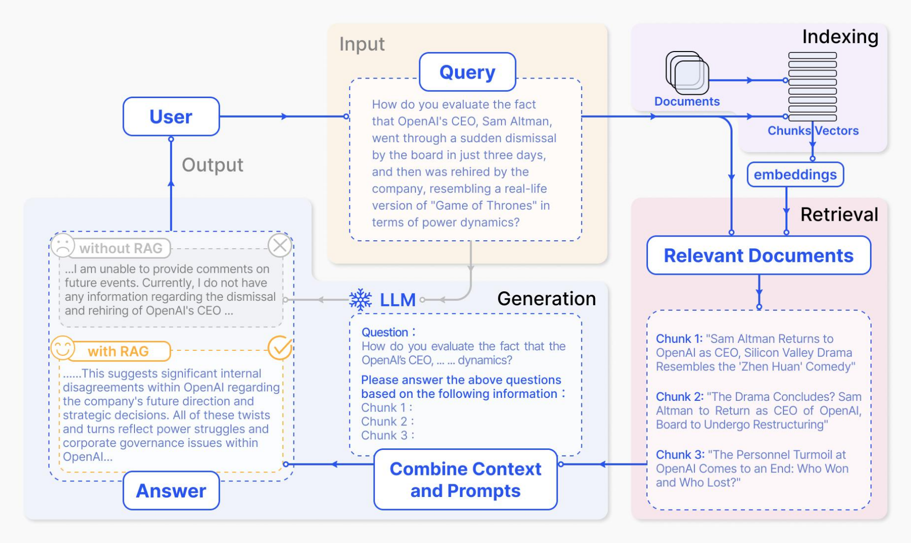

Input
Query
How do you evaluate the fact that OpenAI's CEO, Sam Altman, went through a sudden dismissal by the board in just three days, and then was rehired by the company, resembling a real-life version of "Game of Thrones" in terms of power dynamics?
Indexing
Documents
Chunks
Vectors
embeddings
Retrieval
Relevant Documents
Chunk 1: "Sam Altman Returns to OpenAI as CEO, Silicon Valley Drama Resembles the 'Zhen Huan' Comedy"
Chunk 2: "The Drama Concludes? Sam Altman to Return as CEO of OpenAI, Board to Undergo Restructuring"
Chunk 3: "The Personnel Turmoil at OpenAI Comes to an End: Who Won and Who Lost?"
User
Output
without RAG
...I am unable to provide comments on future events. Currently, I do not have any information regarding the dismissal and rehiring of OpenAI's CEO ...
with RAG
......This suggests significant internal disagreements within OpenAI regarding the company's future direction and strategic decisions. All of these twists and turns reflect power struggles and corporate governance issues within OpenAI...
LLM
Generation
Question：
How do you evaluate the fact that the OpenAI's CEO, ... dynamics?
Please answer the above questions based on the following information：
Chunk 1 :
Chunk 2 :
Chunk 3 :
Combine Context and Prompts
Answer

Fig. 1. An intuitive example of RAG [13].

2.3 Implementation Details

The complete steps of this project are as follows. First, get an OpenAI API Key from the OpenAI platform [14]. After that, open the advanced system properties, and click on "environment variables", and then add a system variable named OPENAI_API_KEY. The value of the variable should be the API Key. The second step involves installing Docker. Subsequently, install Milvus service and launch it. The third step is to install Python3 via Anaconda, then install the requirements of this project. Note that CUDA version pytorch is as it allows the embedding model to process text and generate vectors more rapidly. Finally use python to execute cli.py and import the law data using command build. When the processing to the data is completed, enter the ask command to get the answer. This work introduces several changes, enabling the LLM to provide a response even when uncertain, and to import English law text [15].

3 Results and Discussion

Since RAG is a relatively nascent field, there are many ways to evaluate its performance such as RAG benchmarking [16], and Automated Evaluation of Retrieval Augmented Generation (RAGAS) [17] etc. Among these methods the most popular one is RAGAS. Thus, the method was used to evaluate this model as shown in Fig. 2 and indicates that faithfulness and answer relevancy is the highest of the four standards. It indicates that faithfulness is scored at 0.9444, answer relevancy at 0.7488, while context precision and context recall are measured at 0.4 and 0.45, respectively.

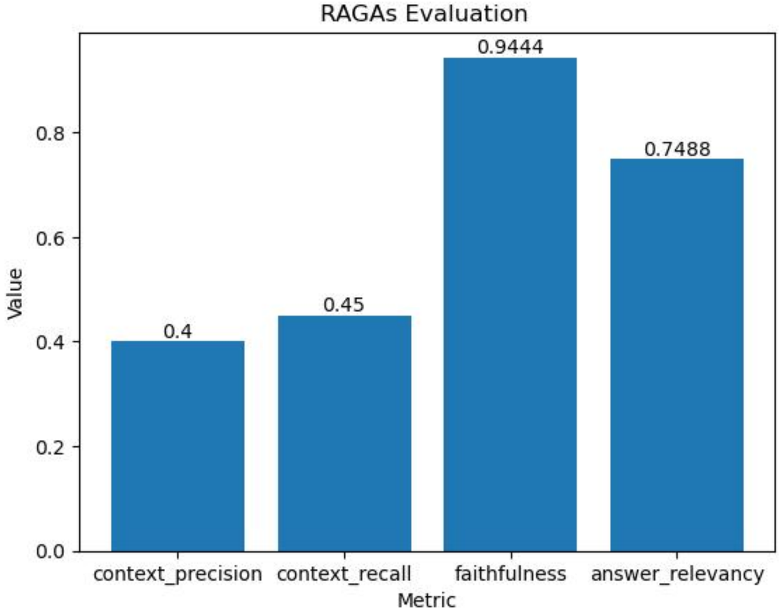
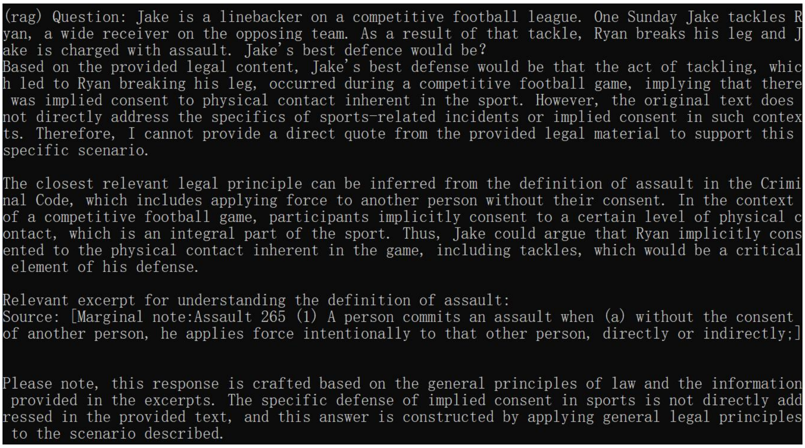


Fig. 2. The evaluation result based on RAGAS (Photo/Picture credit : Original).



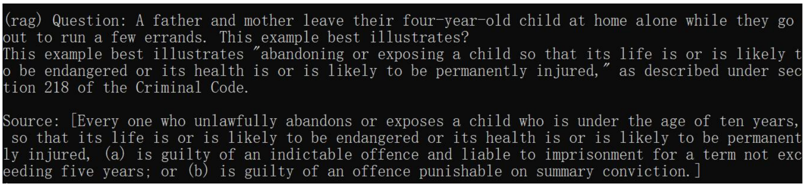
```
(rag) Question: Jake is a linebacker on a competitive football league. One Sunday Jake tackles R
yan, a wide receiver on the opposing team. As a result of that tackle, Ryan breaks his leg and J
ake is charged with assault. Jake's best defence would be?
Based on the provided legal content, Jake's best defense would be that the act of tackling, whic
h led to Ryan breaking his leg, occurred during a competitive football game, implying that there
 was implied consent to physical contact inherent in the sport. However, the original text does
not directly address the specifics of sports-related incidents or implied consent in such contex
ts. Therefore, I cannot provide a direct quote from the provided legal material to support this
specific scenario.

The closest relevant legal principle can be inferred from the definition of assault in the Crimi
nal Code, which includes applying force to another person without their consent. In the context
of a competitive football game, participants implicitly consent to a certain level of physical c
ontact, which is an integral part of the sport. Thus, Jake could argue that Ryan implicitly cons
ented to the physical contact inherent in the game, including tackles, which would be a critical
 element of his defense.

Relevant excerpt for understanding the definition of assault:
Source: [Marginal note:Assault 265 (1) A person commits an assault when (a) without the consent
of another person, he applies force intentionally to that other person, directly or indirectly;]


Please note, this response is crafted based on the general principles of law and the information
 provided in the excerpts. The specific defense of implied consent in sports is not directly add
ressed in the provided text, and this answer is constructed by applying general legal principles
 to the scenario described.
```

Fig. 3. Question about a football player's tackle (Photo/Picture credit : Original).



```
(rag) Question: A father and mother leave their four-year-old child at home alone while they go
out to run a few errands. This example best illustrates?
This example best illustrates "abandoning or exposing a child so that its life is or is likely t
o be endangered or its health is or is likely to be permanently injured," as described under sec
tion 218 of the Criminal Code.

Source: [Every one who unlawfully abandons or exposes a child who is under the age of ten years,
 so that its life is or is likely to be endangered or its health is or is likely to be permanent
ly injured, (a) is guilty of an indictable offence and liable to imprisonment for a term not exc
eeding five years; or (b) is guilty of an offence punishable on summary conviction.]
```

Fig. 4. Question about parents leaving their child at home alone
(Photo/Picture credit : Original).

Almost all the evaluate method acquire a dataset include four parts, questions, answers, contexts and ground truth. Unfortunately, this kind of dataset was not prepared, especially for specific law. Thus, this study has to create an original one from some quiz about Canada criminal law [18]. So finally, 10 questions were used in the dataset and most of them are indirect questions like Mary murdered her roommate, Annie etc. Although these questions are complicated and some are not very suitable to answer directly because they're originally selected questions, the model can deal with it well as shown in Fig 3. The question is not easy and a normal person with Canada criminal code text is not likely to answer it easily and quick, but the RAG model answered it well though it's a little long. However, it still has the disadvantage of a LLM, sometimes it can't fully understand the question as in Fig.4. It considers that parents leaving their child alone at home while they enjoy activities outside as abandonment, though this is not necessarily the case.

As shown in Fig.2, the faithfulness of the model is close to 1, indicating that the model is factually accurate and very reliable. Because the higher this index, the less it will be affected by hallucinations. Answer relevancy shows how relevant is the generated answer to the question. 0.7488 is quite well considering this evaluation only used 10 samples and the questions are not very straightforward. However, the performance of context precision and context recall is relatively low. But this should only be taken as a reference due to the evaluation method not being perfect while law answers can't be so precise as other RAG models. For instance, some questions require detailed answers to be factually correct, but the provided context might be interpreted as noise, thus affecting the context precision and recall ratings.

4 Conclusion

In this work, a LLM solving law problems using RAG is proposed. RAG is used to overcome hallucination and provide reference. The model was trained with Canada criminal code text as its data. Finally, an evaluation dataset was created and used to start a RAGAS evaluation, which indicates that the LLM's hallucination problem is basically solved, and answers are quite relevant to the questions. However, some questions still need to be solved, such as instances where the LLM fails to understand the user's questions and the fact that the evaluation dataset is neither standardized nor sufficiently large. Furthermore, the study is not using a graphics platform so that the users can't use it easily. Therefore, it is hoped that in the future the model will be implemented on a graphical platform with a broader range of laws to assist legal professionals and the general public in their work and daily lives. Meanwhile, efforts should be made to create or find a standardized dataset and method more suitable to evaluate and enhance its performance.

References

1. Liu, Y., Liu, L., Yang, L., Hao, L., Bao, Y.: Measuring distance using ultra-wideband radio technology enhanced by extreme gradient boosting decision tree (XGBoost). Automation in Construction 126, 103678 (2021).
2. Qiu, Y.J., Wang, J.: A Machine Learning Approach to Credit Card Customer Segmentation for Economic Stability. In: Proceedings of the 4th International Conference on Economic Management and Big Data Applications, ICEMBDA 2023, October 27–29, 2023, Tianjin, China (2024).
3. Liu, Y., Bao, Y.: Intelligent monitoring of spatially-distributed cracks using distributed fiber optic sensors assisted by deep learning. Measurement 220, 113418 (2023).
4. Qiu, Y., Hui, Y., Zhao, P., Cai, C.H., Dai, B., Dou, J., Bhattacharya, S., Yu, J.: A novel image expression-driven modeling strategy for coke quality prediction in the smart cokemaking process. Energy 294, 130866 (2024).
5. Zhao, F., et al.: A new method using LLMs for keypoints generation in qualitative data analysis. In: 2023 IEEE Conference on Artificial Intelligence (CAI). IEEE, (2023).
6. Athaluri, S.A., et al.: Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus 15(4), (2023).
7. Shuster, K., et al.: Retrieval augmentation reduces hallucination in conversation. arXiv preprint arXiv:2104.07567 (2021).
8. Louis, A., van Dijck, G., Spanakis, G.: Interpretable long-form legal question answering with retrieval-augmented large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 20, (2024).
9. Chouhan, A., Gertz, M.: LexDrafter: Terminology Drafting for Legislative Documents using Retrieval Augmented Generation. arXiv preprint arXiv:2403.16295 (2024).
10. Department of Justice Canada: Criminal Code (R.S.C., 1985, c. C-46). Department of Justice, 2024, laws.justice.gc.ca/eng/acts/C-46/FullText.html. Accessed.
11. wxywb: history_rag. GitHub, 27 Apr. 2024, github.com/wxywb/history_rag.
12. Liu, Y., et al.: Summary of chatgpt-related research and perspective towards the future of large language models. Meta-Radiology (2023): 100017.
13. Gao, Y., et al.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 (2023).
14. OpenAI: Quickstart. OpenAI, 27 Apr. 2024, platform.openai.com/docs/quickstart?context=curl.
15. Flash201524: Law_Rag. GitHub, 27 Apr. 2024, github.com/flash201524/law_rag.
16. Chen, J., et al.: Benchmarking large language models in retrieval-augmented generation. Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 16, (2024).
17. Es, S., et al.: Ragas: Automated evaluation of retrieval augmented generation. arXiv preprint arXiv:2309.15217 (2023).
18. Oxford University Press: Student Resources for Ruddell's Criminal Law. Oxford University Press Learning Link, 2024, learning-link.oup.com/access/ruddell2e-student-resources#tag_all-chapters. Accessed 3 May 2024.