Avoiding Perjury of Large Language Models using Retrieval Augmentation

Tomás Vergara Browne
Pontificia Universidad Católica de Chile
tomvergara@uc.cl

Abstract

Retrieval Augmentation provides the abilities to Large Language Models to access external information for them to generate more factual answers. Although the use of Retrieval Augmentation techniques has been intensely studied in the last year, a particular domain of legal domains has not been significantly studied, in comparison to other domains. In this paper, we explore the use of Retrieval Augmentation for the Mistral 7B model, which has surprised the open source community for its ability to supersede in performance models which are considerably larger. Through the use of the open source tools in LangChain, we are able to add retrieval capabilities to the model. With this model, we are might be able to a strong evaluation benchmark of LexGLUE. Currently, we are not able to get strong results, particularly for the choice of using a smaller 1.3B model called BLING. Before the final version, we expect to implement the use of Mistral 7B and get much stronger results. We make our code publicly available for reproducibility and auditability of the current development.[1]

1 Introduction

Retrieval Augmented Generation (RAG) is a technique to combine information retrieval, with a text generator. I has been extensively studied over this last year the use of RAG with Large Language Models (LLMs), showing that it can substantially improve a model's performance (Ram et al., 2023). One particular aspect that this technique is allowing to improve is the models factuallity of its answers (Lewis et al., 2020).

Altough there has been some exploration on the use of LLMs in the legal domain (Chalkidis et al., 2020), there has not been that much focus on the use of RAG techniques to enhance a model's abilities to perform on standarized legal benchmarks, such as LexGLUE (Chalkidis et al., 2022). For example, (Savelka et al., 2023) uses a RAG technique with GPT-4 and empirically shows that it makes the model less prone to hallucinations and provides more factual answers when explaining legal concepts.

Also, the model Mistral 7B (Jiang et al., 2023) has arised as a potential state of the art in relatively small open source LLMs. It is able to outperform much bigger models in many domains, such as Llama 2 with 30B (Touvron et al., 2023b).

This calls for the natural approach of combining RAG and Mistral 7B, into the LexGLUE benchmark. Considering the advances that this model and RAG can make in performance, it is not unrealistic to aim for a new state of the art in the LexGLUE benchmark.

2 Related Work

2.1 LLMs in the legal domain

There has been much work into fine tuning models for the legal domain. One particular example is LEGAL-BERT (Chalkidis et al., 2020). This model currently holds the state-of-the-art in the LexGLUE benchmark.

2.2 RAG in the legal domain

There has been some work into the use of RAG techniques in the legal domain. For example, (Savelka et al., 2023) uses a RAG technique with GPT-4 and empirically shows the model provides more factual answers when explaining legal concepts.

3 Methods

We are using the open source framework of LangChain to implement RAG techniques into open source models. Specifically, we are using the ChromaDB tools inside of LangChain to implement a knowledge base in a vector database.

[1] https://github.com/tvergara/RAG-Lawyer

Dataset	ECtHR A	ECtHR B	SCOTUS	EUR-LEX	LEDGAR	UNFAIR-ToS
Model	μ-F1 / m-F1	μ-F1 / m-F1	μ-F1 / m-F1	μ-F1 / m-F1	μ-F1 / m-F1	μ-F1 / m-F1
TFIDF+SVM	62.6 / 48.9	73.0 / 63.8	74.0 / 64.4	63.4 / 47.9	87.0 / 81.4	94.7 / 75.0
BERT	71.2 / 63.6	79.7 / 73.4	68.3 / 58.3	71.4 / 57.2	87.6 / 81.8	95.6 / 81.3
RoBERTa	69.2 / 59.0	77.3 / 68.9	71.6 / 62.0	71.9 / 57.9	87.9 / 82.3	95.2 / 79.2
DeBERTa	70.0 / 60.8	78.8 / 71.0	71.1 / 62.7	72.1 / 57.4	88.2 / 83.1	95.5 / 80.3
Longformer	69.9 / 64.7	79.4 / 71.7	72.9 / 64.0	71.6 / 57.7	88.2 / 83.0	95.5 / 80.9
BigBird	70.0 / 62.9	78.8 / 70.9	72.8 / 62.0	71.5 / 56.8	87.8 / 82.6	95.7 / 81.3
Legal-BERT	70.0 / 64.0	80.4 / 74.7	76.4 / 66.5	72.1 / 57.4	88.2 / 83.0	96.0 / 83.0
CaseLaw-BERT	69.8 / 62.9	78.8 / 70.3	76.6 / 65.9	70.7 / 56.6	88.3 / 83.0	96.0 / 82.3
RAG-Lawyer			**21.9 / 02.9**			

Table 1: Performance comparison of different models on legal datasets.

We are using an embedding based on the model General Text Embeddings (GTE) (Li et al., 2023). This is an open source embedding model which is currently state-of-the-art in language embeddings benchmarks.

As of the moment, the information in the knowledge base is extracted from the Basic Laws book 2016 edition (found here). We expect to add more knowledge sources before the end of the project.

Currently the language model used is an instruction-finetuned version of Llama 1.3B (Touvron et al., 2023a) called BLING (Xia et al., 2023). This model is not very powerful, and we have seen that its results are rather underwhelming. This choice of model was due to the complications in running the Mistral 7B locally, which made local development considerably harder.

We will change the model to Mistral 7B, as soon as we implement the differences to be able to run the code in a GPU environment in the IALab's cluster.

4 Results

As of the moment, the only benchmark inside of LexGLUE that we implemented for evaluation is the US Supreme Court dataset (SCOTUS). It is a single-label multi-class classification task, where given a document (court opinion), the task is to predict the relevant issue areas. The results obtained for the model can be seen in Table 1.

As mentioned before, the model's performance is quite poor, mostly due to the lack of capabilities in Llama 1.3B. This model is quite small in comparison to current LLMs and it is significantly undertrained aswell. Its poor performance is not an indicator of a bad approach, but rather an expected result.

5 Discussion

Although current results are not promising, we expect to see a significant improvement in performance once we are able to run the Mistral 7B model. We expect to see a strong evaluation in the LexGLUE benchmark, and we expect to see a significant improvement in the model's factuallity.

We also need to write the scripts for each of the individual benchmarks in the LexGLUE benchmark. Due to the fact that the current model took more than 4 hours to complete the evaluation in the SCOTUS dataset, and considering that we need to use the Mistral 7B model, we are prioritizing the implementation of GPU use of the model, and parallel evaluation for several test cases, before actually implementing the other benchmarks.

One disclaimer is that although we claimed that this might lead to a new state-of-the-art on the benchmark, this was highly misleading. The current evaluations on the benchmark were made exclusively with really small finetuned models. It is not a fair comparison to use an LLM as Mistral 7B for this dataset. If we would consider that to be fair, then using GPT-4 should also be fair. It would be unreasonable to expect that the Mistral 7B would be able to win against GPT-4, even with RAG techniques.

Also, previously we proposed using the RETA-LLM framework (Liu et al., 2023). We found that this framework was not as easy to use as the authors claimed. We found a much better experience in implementing RAG techniques using LangChain, which is the current standard for RAG in LLMs.

References

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Jiongnan Liu, Jiajie Jin, Zihan Wang, Jiehan Cheng, Zhicheng Dou, and Ji-Rong Wen. 2023. Reta-llm: A retrieval-augmented large language model toolkit. *arXiv preprint arXiv:2306.05212*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00082*.

Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. 2023. Explaining legal concepts with augmented large language models (gpt-4). *arXiv preprint arXiv:2306.09525*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. 2023. Sheared llama: Accelerating language model pre-training via structured pruning. *arXiv preprint arXiv:2310.06694*.