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 framework of RETA-LLM, we are able to add retrieval capabilities to the model. With this model, we are able to provide a new state of the art for the benchmark of LexGLUE.

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 (**?**). One particular aspect that this technique is allowing to improve is the models factuallity of its answers (**?**).

Altough there has been some exploration on the use of LLMs in the legal domain (**?**), 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 (**?**). For example, (**?**) 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 (**?**) 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 (**?**).

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 Fine tuning large language models in the legal domain

2.2 Retrieval augmented large language models in the legal domain

3 Methods

We are using the open source framework of RETA-LLM (**?**) to enhance Mistral 7B with RAG capacities. This framework allows as to do a "*plug-and-play*" approach, in which there is no need for fine-tuning the model to know how to use the external memory.

For our external knowledge base, we are going to use data from the US and European Law, given that the LexGLUE benchmark is based in legal cases from the US and countries within Europe.

4 Results

5 Discussion

6 Appendices