MSc Project - Reflective Essay

Project Title:	From Fine-Tuning to Retrieval-Augmented Generation: Enhancing Document Question Answering with Large Language Models
Student Name:	**Jay Maheshbhai Agrawal**
Student Number:	**230913640**
Supervisor Name:	Dr. Paulo Oliva
Programme of Study:	MSc. Computer Science

Overview:

It has been both a challenging and learning experience to work on creating the document chat application with the help of large language models. The goal of this project was to take advantage of the possibilities of large language models (LLMs) for processing documents which are usually long and cannot fit into the LLM model's input limit. The shift from traditional fine-tuning methods to a Retrieval-Augmented Generation (RAG) system marked a significant turning point, addressing many of the initial challenges encountered.

In this reflective essay, the strengths and weaknesses of the project are discussed, potential future works are proposed, and theoretical knowledge and its practical application are examined. Also it discusses the legal, social, ethical and sustainability implications of implementing a such system.

Strengths:

One of the main strengths of the project was the use of a diverse legal dataset in the fine-tuning phase of the project. I used the CUAD dataset available on Huggingface. In this dataset, legal documents, contracts, agreements and many other types were incorporated so that the models would have a broad foundation for learning. By using a diverse set of legal questions and queries, the models were trained to recognize different legal terminologies and structures which help in generalizing the models and making them more robust in the real-world search or query scenarios.

Another strength of the project was leveraging the use of advanced NLP Large Language Models like BERT (Bidirectional Encoder Representations from Transformers) and LLaMA (Large Language Model Meta AI) models. While BERT had been found to be proficient in context and semantics, LLaMA was found to be capable of handling longer input sequences effectively. It made possible to investigate the models deeper, evaluate their performance and reveal their advantages and possible weaknesses in the context of legal documents' analysis.

The fine-tuning of the LLaMA model was done effectively using the Low-Rank Adaptation (LoRA) Parameter-Efficient Fine-Tuning (PEFT), which is also a strength for the project. This approach allowed for the efficient adaptation of the large LLaMA model without the need for full retraining, significantly reducing computational costs and resource requirements. This is attained by including a low-rank feature of the weights update wherein LoRA is capable of identifying vital model shift with minimal parameters. Besides, this method enlarged the applicability of the fine-tuning process and helped the project achieve faster progress and try different variants by using several GPUs at a time, thus increasing the effectiveness of the project in working on legal documents. The example of LoRA shows the successful implementation of a new way of using large models efficiently and that in detail contributes to the success of the project in terms of employing more sophisticated LLMs.

Finally, testing our whole methodology ensured that my project trajectory was based on results and facts that I recorded. The testing included the fine-tuned models, RAG system where as well the general testing of the application both on the frontend and backend side. Fine-tuned models were tested using a SQuAD metrics which shows the F1 and Exact match scores of the model. The fine-tuned BERT model showed an F1 score of 56.54 and LLaMA with F1 score of 20.23. RAG system was tested with several different configurations and each of them were evaluated based on their accuracy. Finally, the most efficient configuration was selected for our RAG system. This way of testing provided confidence in the system and the ability to answer for different form of query from the users. The above evaluation procedure ensured that most of the problems that could have affected the system were corrected thus improving on its stability and performance.

Another strength of my project was implementing an effective RAG (Retrieval Augmented Generation) system. The development and incorporation of the RAG system successfully eliminated the hallucination problems that came with the fine-tuned LLaMA model. Altogether, the integration of retrieval mechanisms and generative capabilities made the use of RAG guarantee that responses are factual and not just random or unrelated information, which improved the effectiveness of the system massively. This implementation solved important concerns in the project which were context length and the hallucination issues.

Weaknesses:

Using LLMs for fine-tuning and generative purposes requires GPU resources. Fortunately, I was able to use Apocrita's HPC cluster to perform GPU intensive tasks. I was only able to use the cluster for 1 hour window and after that I had to wait in the queue again to get access to the cluster. Due to this, I implemented checkpointing approach to save fine-tuning process at several steps and continue again from the latest checkpoint in the next 1 hour window.

One of the limitation of my project was a large amount of computational power needed for, first, running backend server with large language models and, second, fine-tuning large language models, such as LLaMA. The high demand for processing power and memory limited the system's scalability and accessibility, potentially restricting its deployment in resource-constrained environments. This challenge highlights the need for optimizing resource usage and exploring more efficient model architectures or methods in future work. Due to this, I had to quantize the LLMs to 4bit which helps in utilizing the memory efficiently but it degrades the quality of responses produced.

Another weakness was LLaMA model showing poor performance on the unseen data. Even though the fine-tuned LLaMA model has a more complex architecture, it shows worse performance than BERT, mainly due to hallucinations and memorizing the training data. Despite the fact that LLaMA was supposed to be proficient in dealing with longer contexts, the creation of unrelated or wrong information became evident with it. This strongly highlights the fact that fine-tuning ends up being a difficult task in large models and the need to address the hallucination challenges that models present when maximizing their potential.

Possibilities of Future Work:

Some of the areas that can be further worked on in the project are as follows which in a way would enhance the current capabilities and also expand its usage: One of such areas, which can be further enhanced, is the usage of feedbacks and personalization components. This way the users have the capability to feedback for the system and in return respond to interactively generated responds whereby the system will able to change the outputs based on the needs of the user in as much as it finds suitable response to other inputs. A moderate level of personalization could be where some content is pushed towards the user including response history and feedback history which are captured in advance to tailor the response to better fit the user.

The second interesting avenue for future research is to supplement the information stated concerning ethical problems and their means of reducing bias. While language models become more integrated to the decision made, it is important to develop methods in which bias in the answers generated can be detected. It may include developing models that can be able to identify and remove biased words or act on the content and consolidate the information that the models used for training came from balanced data. This implies that the system has to bring quality data and at the same time be non-biased when presenting the findings.

Work That Could Have Been Conducted with More Time:

With more time, I would try to reduce hallucinations in finetuning the LLaMA model. I could have fine-tuned all hyperparameters of learning rates, batch sizes, epochs, and LoRA which possibly might help in minimizing these problems. Changing the learning rate may effectively mitigate overfitting since the model will learn new information without over-training. If the batch size and epoch values are altered, this would give a picture of the model’s training status. Additionally, researching different possibilities of LoRA might improve the stability of the model in the context of legal text understanding. Such an approach would have helped to get better results concerning the model’s performance.

Besides the fine-tuning, more time would have helped in the research of the dynamic retrieval mechanisms in order to improve the RAG system. Dynamic retrieval means changing the method of retrieval according to the nature and toughness of the user queries and it has the potential of making the document chunks more relevant and precise. This research could result in advanced RAG system that would offer the best information, increasing the efficiency and effectiveness of the document chat application.

More time would have been devoted to the extent to add support to more kinds of documents and the enhancement of the user interface. For example, the facilities such as the document formats like the .docx and the .html among others were helpful to the system in as much as addressing the diverse user demand. Additionally redesigning and implementing the user interface and the use of functionalities such as real time response, customize, and analyze would help in the enhancement of the interaction between the user and the system.

Critical analysis of the relationship between theory and practical work produced:

It is clear from the project that the theory and practice are aligned in a very complex manner but easily understandable. In theory, RAG systems are a promising solution because of the possibility to combine retrieval-based and generative components, always providing solid facts. This theoretical framework was practically applied when adding a retrieval system based on FAISS and a re-ranking using the CrossEncoder model. In this case therefore the implementation showed practicality where theoretical studies were in agreement and showed that RAG indeed helped to decrease hallucination and increase response accuracy when analyzing large documents.

But with reference to the fine-tuning of the LLaMA model, it was somehow difficult to interpret the relationship between the theory and the practice. In theory, using approaches like LoRA (Low-Rank Adaptation) to fine-tune very large language models should enable efficient repurposing of the models for domain-specific applications without much computing power. In Real-World setting, although LoRA was making it efficient to fine-tune the models, the actual gains in terms of performance improvements were not quite as expected notably when it came to hallucinations. This was mainly due to the fact that theoretically derived fine-tuning approach could not simply be scaled up to handle large models and, as a result, the study brought out the possible additional need for new training parameters experimentation to realize such goals.

Last of all, the project also highlighted the interaction between the theoretical framework and the practical implementation of the system architecture. In theory, the architecture based on Flask on the backend, and React.js at the frontend will allow for the application to be scalable and flexible in terms of deployment. In practice, this theoretical framework was successfully applied; thereby creating a reliable system which can process user interactions as well as large quantities of data. The application of theoretical design principles in the system architecture also helped in the integration process of the system components and the success of the project.

Awareness of Legal, Social and Ethical Issues and Sustainability:

From a legal point of view the work complies with all legal requirements including those related to the use of licensed models and datasets. BERT and LLaMA LLMs were used under the right license to prevent violation of copyrights and other similar rights. Further, data management practices in the system integrated measures to protect users' data and, where any sensitive data was to be captured, it was ensured that the handling of such data was well protected.

One of the primary ethical challenges was to provide accurate and reliable of the information. The project emphasized the importance of generating factual and accurate responses, particularly in contexts where users rely on the system for decision-making. To exclude certain hallucinatory answers, the system also implemented special features for improving the relevance of the context; this way, the number of factual inaccuracies was minimized.

Another important factor that was given much consideration over the lifecycle of the project was sustainability, especially about the carbon emissions by LLMs. The computational demands of training and deploying LLMs, such as BERT and LLaMA, can have significant environmental impacts due to high energy consumption. To address this, the project focused on optimizing computational efficiency and minimizing resource usage. This involved selecting efficient models and configurations, implementing best practices for energy-efficient computing, and leveraging cloud-based infrastructure to reduce the carbon footprint. By using scalable computing resources and optimizing model parameters, the project minimized energy consumption while maintaining high performance.