Working title: Advancements and Applications of Large Language Models: A Comparative and Contextual Analysis

Overall aim: The overarching goal of this thesis is to explore the efficacy of Large Language Models (LLMs), particularly the advanced Mistral-8x7b, in the realm of scientific disciplines such as physics, chemistry, and biology. This exploration entails a meticulous fine-tuning of the LLM using the SciQ dataset, a specialized resource in scientific question-answering, to achieve superior performance in these specific fields. The study aims to demonstrate how targeted fine-tuning with domain-specific datasets can significantly enhance an LLM's understanding and handling of complex scientific queries and data. This process will not only test the adaptability and learning capabilities of the Mistral-8x7b model but also seek to establish a benchmark for LLM performance in processing and interpreting scientific information, thereby contributing valuable insights into the application of LLMs in specialized academic and research-based scenarios. This entails a thorough comparison with the LLaMa-2-13b model over a variety of datasets to evaluate performance measures such as accuracy and computing efficiency. Furthermore, the study attempts to analyse the practical usability of these models in certain business contexts, as well as the reasoning behind picking selected datasets for LLM applications.

Objectives:

1) Performance Improving: To fine-tune the Mistral-8x7b model using the Sciq dataset, aiming to significantly improve its performance in answering questions related to physics, chemistry, and biology.
2) Comparative Analysis: Conduct a comparative study of the Mistral-8x7b model's performance against the LLaMa-2-13b model in the context of scientific datasets, focusing on accuracy and computational efficiency.
3) Model Optimization Techniques: Implement and evaluate the effectiveness of Low Rank Adaptation (LoRA) and Parameter Efficient Fine Tuning (PEFT) techniques to fine-tune the Mistral-8x7b model for domain-specific tasks.
4) Performance Metrics Evaluation: Assessing the improvements in the fine-tuned model using relevant metrics, such as Bleu, Rouge, Perplexity metrics, highlighting advancements in answering complex scientific queries.
5) Implementation Showcase: Pushing the final model on HuggingFace platforms and creating a space with Gradio, highlighting the application of the developed LLM in chosen business scenarios.

Methodology:

- Model Selection
    - Choosing the Mistral-8x7b and LLaMa-2-13b models for a comprehensive comparative analysis.
    - Prioritizing the Mistral-8x7b for fine-tuning due to its advanced architecture and potential for optimization.
- Dataset Selection
    - Employing the Sciq dataset, which contains questions and answers in physics, chemistry, and biology, to fine-tune and test the model.
- Fine-Tuning Techniques
    - Implementing Low Rank Adaptation (LoRA) and Parameter Efficient Fine Tuning (PEFT) to optimize the Mistral-8x7b model.
    - Applying these techniques to enhance the model's performance specifically in the context of the Sciq dataset.
- Comparative Performance Analysis
    - Conducting tests using both the Mistral-8x7b and LLaMa-2-13b models to assess and compare their performance.
    - Conducting tests using the Mistral-8x7b model before and after fine-tuning and compare their performance.
    - Focus on performance metrics.

- Cost and Resource Analysis
    - Estimating the overall costs involved in the fine-tuning process, including computational resources and time.
- Documentation
    - Ensuring thorough documentation of the methodology for reproducibility.
    - Including detailed descriptions of the fine-tuning processes, parameter settings, and evaluation protocols.

Structure:

The thesis commences with an Introduction, outlining the significance of fine-tuning of LLMs with modern techniques in text and defining the research objectives, scope, and application fields.

The next section, Theoretical Background, reviews essential concepts in machine learning and neural networks, progressing to advanced topics like CNNs, RNNs, LSTMs, and Transformers. This foundation is crucial for understanding the later practical applications in LLMs.

The LLMs Overview follows, discussing its evolution and methodologies, setting the stage for fine-tuning techniques. In the LLMs section, the focus Low Rank Adaptation (LoRA) and Parameter Efficient Fine Tuning (PEFT) approaches, their importance, challenges, and current research trends.

The Methodology section details the research design, encompassing data collection, preprocessing, feature extraction, fine-tuning, model training, and model evaluation criteria.

Data Analysis and Findings presents and compares the results of various models, linking these findings back to the theoretical concepts.

This study will research of comparison of the Mistral-8x7b and LLaMa-2-13b models' performance. Then it will be processed of fine-tuning the Mistral-8x7b model using the Sciq dataset. Evaluation of the improvements and effectiveness of the fine-tuning techniques will be observed.

Finally, the thesis concludes with Conclusion and Future Work, summarizing key insights, reflecting on objectives, and suggesting future research directions.

Preliminary reading list:

1) Eliseev, A., & Mazur, D. (2023). *Fast Inference of Mixture-of-Experts Language Models with Offloading* (arXiv:2312.17238). arXiv. http://arxiv.org/abs/2312.17238
2) Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. de las, Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., … Sayed, W. E. (2024). *Mixtral of Experts* (arXiv:2401.04088). arXiv. http://arxiv.org/abs/2401.04088
3) Lu, J., Yu, L., Li, X., Yang, L., & Zuo, C. (2023). LLaMA-Reviewer: Advancing Code Review Automation with Large Language Models through Parameter-Efficient Fine-Tuning. *2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE)*, 647–658. https://doi.org/10.1109/ISSRE59848.2023.00026
4) Wan, Z., Wang, X., Liu, C., Alam, S., Zheng, Y., Liu, J., Qu, Z., Yan, S., Zhu, Y., Zhang, Q., Chowdhury, M., & Zhang, M. (2023). *Efficient Large Language Models: A Survey* (arXiv:2312.03863). arXiv. http://arxiv.org/abs/2312.03863