MNLP Project Proposal

Arthur Chansel | 324265 | arthur.chansel@epfl.ch
Iris Segard | 310860 | iris.segard@epfl.ch
Gilles de Waha | 330072 | gilles.dewaha-baillonville@epfl.ch
MoreNectarLessPoison

1 Introduction

This project proposes to build a real-world language model (LM) for educational assistance. In particular it will be answering scientific questions from academic courses. Our implementation is based on LLaMa language models (Touvron et al., 2023), using the open source pretrained model Llama 3 Instruct. We are presenting several methods to get the most performing chat bot.

2 Model

2.1 Generator Model

We aim to utilise LLaMa 3 as our initial model which we would then fine-tune with the help of DPO (Rafailov et al., 2023) and it's variants (such as ODPO (Amini et al., 2024) and IPO (Azar et al., 2023)). We have also researched backup options to explore if needed.

LLaMa 3 : LLaMa 3 is a family of autoregressive language models by Meta available in two sizes : 8B and 70B parameters, each with a pretrained and instruction-tuned variant. Due to our limited computational capacity, we will use the 8B parameter variant. As we strive to create an educational chat-bot, we will use the Instruct variant as it is already aligned through SFT and RLHF for helpfulness and safety.

However, even when selecting the 8B parameter model, fine-tuning every weight of the model is extremely resource and data intensive.

To remedy this, we propose using an adapter such as LoRa (Hu et al., 2021) instead. We would therefore freeze all 8B parameters from our foundation model and inject new learnable parameters in the model to learn. It should allow us to keep computational costs quite low whilst maintaining training efficacy.

Phi 2: Should LLaMa 3 not be suited for our needs, we intend to use Microsoft's model Phi-2 (Mojan Javaheripi, 2023) which has 2.7B parameters. The "out-of-the-box" model is a base model that has not been aligned to human preference or instruct fine-tuned. A Phi-2 Instruct model[1], fine-tuned on the (filtered) ultrachat200k dataset[2], is however available on Hugging Face.

2.2 RAG and MCQ Specialization

We will augment our model with RAG (Retrieval Augmented Generation). This method helps generating better answers which are knowledge intensive (Patrick Lewis, 2021) by taking in a set of documents alongside the user prompt.

Retrieve: Given the prompt, we will retrieve passages from an external dataset with the RagRetriever from the transformers library [3] and the pretrained facebook-dpr-ctx_encoder-single-nq-base Dense Passage Retrieval model (Karpukhin and et al., 2020).

Augment: The retriever and our generator model will be encapsulated in a RagModel from the transformers library which prepends the retrieved passages to the input and passes them to the generator.

To correctly format the outputs of our model to MCQ questions so that it generates a single letter, we will try two approaches. The first one is simply prompt engineering by asking the model to correctly format the output. And the second one is using the Outlines library which provides a framework to guide text generation (Willard and Louf, 2023).

3 Data

Each student involved in the project labelled a preference dataset, which constitutes our primary data for fine-tuning the model. Each data point consists

[1] https://huggingface.co/venkycs/phi-2-instruct

[2] https://huggingface.co/datasets/HuggingFaceH4/ultrachat_200k

[3] https://huggingface.co/docs/transformers/model_doc/rag

of two generated answers, among which one is preferred. The generator model answers open questions, while the improved model answers MCQ questions. We intend to split the students dataset into these two cases.

We will enhance our data with prompt engineering and Zero-Shot approach (Kojima et al., 2023), to extract the most of our model capabilities.

3.1 Generator Model

Llama 3 Instruct was pretrained over 15 trillion tokens of data from publicly available sources[4], detailed in LLaMa paper (Touvron et al., 2023). Notably, part of its training set is questions and answers from Stack Exchange websites.

We will use more data to fine-tune the model into a desired specific behavior. Eventually, we would complete our data with external sources: Stanford Question Answering (Rajpurkar et al., 2016), HC3[5], ScienceQA[6], CommonsenseQA (Talmor et al., 2019) and OpenBookQA (Mihaylov et al., 2018). However, these are popular public datasets on which LLaMa could possibly already have been trained on.

3.2 RAG Specialization

For the Retrieval Augmented Generation, we will use the default wiki_dpr dataset [7] that contains 21M passages from Wikipedia along with their DPR embeddings.

4 Evaluation

The approach to evaluate our model is discussed below.

Moreover, popular benchmarks may be useful to compare our results, such as GLUE (Wang et al., 2019) and SuperGLUE (Wang et al., 2020).

4.1 Generator Model

For generator model evaluation, we plan to use BertScore (Zhang et al., 2020) and COMET (Rei et al., 2020) metrics. They are reliable methods to assess text generation quality. They leverage embedding space to measure the similarity and coherence of responses. These factors are crucial when answering to scientific questions, where logic is fundamental. Thus, they provide robust measures, and ensure our model's outputs are relevant.

We would like to explore other metrics to assess generation validity in an online approach. We might use HellaSwag (Zellers et al., 2019), which evaluates the completeness of a sentence, to verify that we generate complete and synthetically correct answers. TruthfulQA (Lin et al., 2022) measures truthfulness of answers, it might be useful to verify the model logic.

4.2 RAG and MCQ Specialization

The generations of our model with the RAG specialization will first simply be compared with the generations from our base model according to the above metrics. Then, we will evaluate our retriever by qualitatively determine the relevance of the retrieved passages according to the prompt.

The improved model is specialized to answer to MCQ questions. Thus a common appropriate metric in this case is confusion matrix (Heydarian et al., 2022).

5 Ethics

In the context of our project : an educational chat-bot geared towards engineering university students, we face multiple ethical challenges:

- Our main challenge is that LLM answers are based on the most statistically likely outcome to a prompt. The model is therefore not "obligated" to produce factual answers which can lead to hallucinations and disinformation. In the context of an educational chat-bot, this is a serious concern as it is at odds with the original intent of the model: education. We aim to limit this issue using RAG.

- Another issue stems from the data that it is trained on. Indeed, these datasets are created by humans which inevitably leads to human biases being embedded in the data itself from which the model will learn. It will therefore "imitate" these biases and even amplify them (Ahn et al., 2022). This challenge can be tackled through human alignment during training, however it is not enough. Biases are still present in the model and reveal themselves insidiously.

- Lastly, models can easily become dangerous when vulnerable individuals (such as children or people struggling with mental health issues) interact with them. We consider that our chat-bot will be used for it's intended purpose which makes this issue out of scope for our project.

[4] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[5] https://huggingface.co/datasets/Hello-SimpleAI/HC3

[6] https://scienceqa.github.io/

[7] https://huggingface.co/datasets/wiki_dpr

References

Jaimeen Ahn, Hwaran Lee, Jinhwa Kim, and Alice Oh. 2022. Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of DistilBERT. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 266–272, Seattle, Washington. Association for Computational Linguistics.

Afra Amini, Tim Vieira, and Ryan Cotterell. 2024. Direct preference optimization with an offset.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences.

Mohammadreza Heydarian, Thomas E. Doyle, and Reza Samavi. 2022. Mlcm: Multi-label confusion matrix. *IEEE Access*, 10:19083–19095.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Vladimir Karpukhin and Barlas Oğuz et al. 2020. Dense passage retrieval for open-domain question answering.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Sébastien Bubeck Mojan Javaheripi. 2023. Phi-2: The surprising power of small language models.

Ethan Perez et al. Patrick Lewis. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge.

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. 2023. Llama: Open and efficient foundation language models.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. Superglue: A stickier benchmark for general-purpose language understanding systems.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding.

Brandon T. Willard and Rémi Louf. 2023. Efficient guided generation for large language models.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence?

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.