587 Final: Fine-Tuning LLMs For Research

Ronak Singh

April 25, 2025

1 Introduction

One of the most crucial parts of research is the practice of developing an approach to a particular research question using the scientific method. This includes developing a sound approach which addresses how data is collected, what hypotheses are tested, and how experiments and analysis are conducted. In this project, I fine-tune Meta's open-source Llama-3.2-3B-Instruct model with the objective of assisting in this process, namely generating a feasible approach to a given research problem. In order to fine-tune this model, I created a synthetic dataset of research questions and sample research approaches spanning multiple disciplines using OpenAI's GPT 4.1-mini model. We begin by describing the construction of the synthetic dataset. We continue by discussing details concerning the Llama-3.2-3B-Instruct model, as well as training and evaluation details. Next, we discuss the experiments and provide initial results of the fine-tuned model (using ROUGE and METEOR scores). We end our discussion by analyzing the results and reviewing the lessons and issues I came across during this project.

2 Dataset Construction

To effectively train a model to generate thorough and feasible approaches to a wide variety of research problems, a substantial and representative dataset was required. Due to a lack of an existing such dataset, I opted to generate a synthetic dataset. The approach I took to generate the dataset was to first outline 8 broad *fields* of research: arts, engineering, formal sciences, humanities, life sciences, natural sciences, professional, and social sciences. Next, for each field, I generated a list of *subdomains* within each field. For example, for the formal sciences, the subdomains included mathematics, statistics, computer science, logic, and data science. Across all 8 fields, there are a total of 61 subdomains (see `generate_data.ipynb` to view all subdomains). The reason for having such a diverse set of research fields represented in the dataset is to prevent data imbalances and minimize biases. Next, Open AI's GPT 4.1-mini model was prompted to generate 200 unique research questions for each subdomain. After cleaning the questions to remove any duplicates or similar questions (after cleaning there were over $15,000$ questions), GPT 4.1-mini was prompted to generate research approaches for each of these questions. The model was prompted to pay special attention to hypothesis, proposed methodology, key steps involved, data collection plan (if applicable), data analysis plan (if applicable), potential challenges or limitations, and ethical considerations (if applicable).

An alternative approach that was initially assessed was to use the arXiv API to extract research approaches from real papers to generate the dataset. The primary shortcomings of this approach included the difficulty of extracting a consistent and high-quality summary of each paper's research approach. Oftentimes the summaries would be incomplete or too generalized. My rationale for instead using GPT 4.1-mini to synthetically generate research approaches is that OpenAI's models are likely already trained on a large corpus of research papers, meaning that it is reasonable to assume that any research approaches generated by the GPT 4.1 family of models are likely to be (at least partially) aligned with approaches from existing research papers.

3 LLM Selection

After evaluating several candidate open-source models against the requirements of this project and hardware constraints, Meta's Llama-3.2-3B-Instruct model (available as meta-llama/Llama-3.2-3B-

Instruct on Hugging Face) was selected as the base model for fine-tuning. Llama-3.2-3B-Instruct was selected due to it's strong performance as a base model despite it's relatively small number of parameters. A 3 billion parameter model was selected primarily to allow training to be completed within 6 hours on a single rented GPU. This 6 hour time constraint was set since the service from which I rent the GPU (Paperspace) only allows for GPU resources to be used for 6 hours at a time.

3.1 Training Details

Following the generation and cleaning of the synthetic dataset and the selection of Llama-3.2-3B-Instruct as the base model, The next phase was to train (or fine-tune) the base model on the synthetic dataset. Before using the data for fine-tuning, the dataset was divided into a 75/25 training/validation set, meaning that 9148 question/approach pairs were ultimately used to fine-tune the base model.

Fine-tuning itself was performed primarily using Hugging Face's PEFT (Parameter-Efficient Fine-Tuning) and Supervised Fine-Tuning libraries. I applied LoRA (Low-Rank Adaptation) fine-tuning to all the projection layers in the base Llama-3.2-3B-Instruct model (as is common practice for fine-tuning Llama models). The LoRA attention dimension was set to 16, and the dropout for LoRA layers was set to 0.05. I set the learning rate to $2 * 10^{-4}$ (and applied cosine annealing) and the effective batch size to 8. The model was trained for 2 epochs (see `train.ipynb` to review the training pipeline). All training was performed on a rented NVIDIA A4000 GPU.

4 Evaluation and Experiments

After fine-tuning the base Llama-3.2-3B-Instruct model on the synthetic training dataset, we begin by discussing relevant evaluation metrics to assess the quality of the fine-tuned model's performance. The purpose of these evaluation metrics is to assess how well-aligned the fine-tuned model's generated research approaches are with the existing approaches generated by GPT 4.1-mini. Later, we continue the discussion by discussing the design and results of my experiments with the model (once again, the objective of these experiments is to understand how well the fine-tuned model performs at the target task).

4.1 Evaluation Metrics

To quantitatively assess the quality of responses generated by the fine-tuned model, I utilized two widely recognized evaluation metrics often used for NLP tasks: ROUGE and METEOR. Both evaluation metrics compare a *candidate* text to a *reference* text, and they output a score between 0 and 1, indicating how similar the two texts are (the higher the score, the better). ROUGE primarily focuses on lexical overlap, while METEOR considers synonyms, stemming, and word order when assessing overlap between the reference and candidate texts. Both metrics can provide valuable information about the performance and alignment of the fine-tuned model, assessing both lexical and content-based similarities.

4.2 Experimental Design

In this section, I present the experiment I conducted to assess the quality of my fine-tuned model. The primary objective of the experiment is to compare the responses of the base Llama-3.2-3B-Instruct model and the fine-tuned model in the task of generating an approach to a given research question.

First, the dataset of questions used to evaluate the models was created by selecting 2 questions for each of the 61 subdomains from the validation dataset. This resulted in a dataset of 122 question/approach pairs. Then, the base Llama-3.2-3B-Instruct model and the fine-tuned model were prompted with the questions from each of these pairs, and their respective responses were recorded. Finally, the ROUGE and METEOR score for each model is calculated for each field. In this case, the candidate text is the respective model's response, and the reference text is the research approach from the dataset (originally generated by GPT 4.1-mini). For this experiment, I used the ROUGE-L variant, which measures the Longest Common Subsequence (LCS) between the candidate and reference texts. It is important to note that LCS identifies the longest sequence of words that appear in both texts in the same order, *not necessarily contiguously.* The ROUGE-L metric was used to better-understand

sentence-level similarities between the candidate and reference texts, as word-level similarities are less important for research approaches.

The primary reason for reporting both the ROUGE and METEOR evaluation metrics is the difference between the specific kind of similarity that each metric measures. For example, ROUGE assesses the "surface-level" similarities between two texts, making it useful to identify lexical characteristics like how a prompt is structured (and the use of Markdown). On the other hand, METEOR is far better in assessing the similarity between the meaning of two texts, making it useful to identify the *quality of the content* of a model's response.

4.3 Results and Analysis

As shown in Fig. 1, it is evident that the ROUGE score indicates that the fine-tuned model is better at generating approaches to research questions compared to the base model. In particular, the greatest improvements over the base model came in questions from arts, engineering, life sciences, and natural sciences. The remaining 4 fields saw minor improvements. One interesting finding is that the fine-tuned model appears to have improved in knowledge in the life sciences and natural sciences compared to the other fields, despite the balanced dataset. This could indicate biases or inconsistencies in the underlying dataset used to "pre-train" Llama-3.2-3B-Instruct.

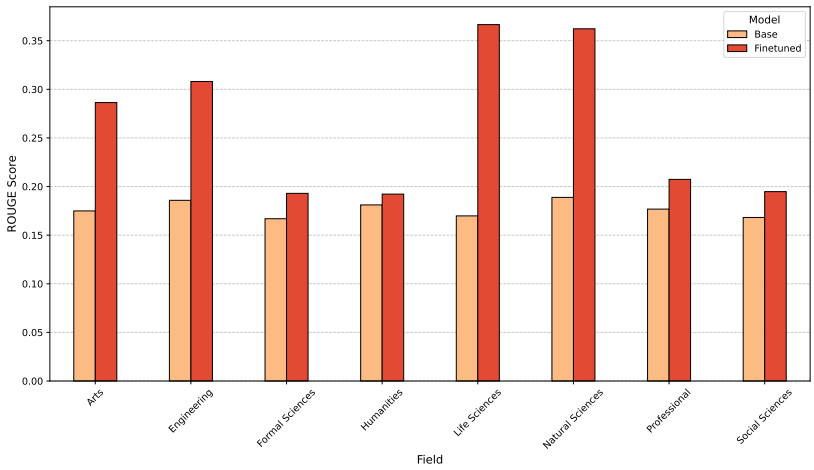

Figure 1: The average ROUGE-L scores of the base Llama-3.2-3B-Instruct model and the fine-tuned model for each of the 8 fields. According to the ROUGE-L data, on average the fine-tuned model outperformed the base model by over 40%.

As shown in Fig. 2, it is once again evident that the METEOR score indicates that the fine-tuned model is stronger in generating research approaches compared to the base model. Interestingly, the fine-tuned model appears to produce significant improvements across all fields other than the humanities (again, despite the balanced dataset). This could indicate that while the formatting of the fine-tuned model's responses isn't well-aligned with the format present in the dataset, the actual content of the model's responses is significantly more aligned with the generated dataset.

5 Conclusions and Personal Thoughts

In summary, based on the conducted experiments and my synthetic dataset, it appears that the fine-tuned model significantly outperforms the base Llama-3.2-3B-Instruct model (by $40\% - 60\%$). While the task of generating approaches to research questions can be applied directly in academic settings, variants of the task can be applied in fields beyond academia. For example, the task of generating approaches to research questions can be adapted to a planning task, where an LLM must generate a feasible plan given a short and noisy objective. These kinds of tasks are being applied today for agentic workflows, as well as for robotics.

Additionally, given the use of synthetic data for fine-tuning, this project can be seen as a short demonstration of distilling GPT 4.1-mini's knowledge into Llama-3.2-3B-Instruct for a particular task

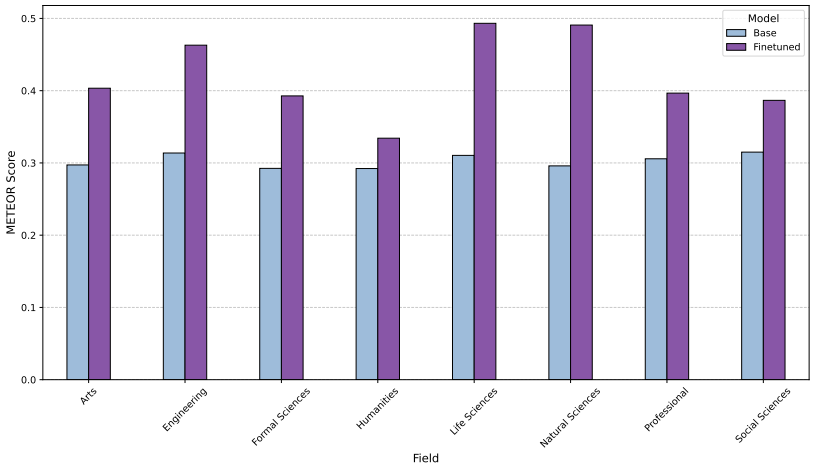

Figure 2: The average METEOR scores of the base Llama-3.2-3B-Instruct model and the fine-tuned model for each of the 8 fields. According to the METEOR data, on average the fine-tuned model outperformed the base model by over 60%.

(generating research approaches). This is significant since we have shown that a larger model like GPT 4.1-mini (which likely requires datacenter-level resources to run) can be distilled to smaller models like Llama-3.2-3B-Instruct (which can potentially run on the edge or with limited GPU resources) with relative success (considering the small dataset and short training time used in fine-tuning). Increasing training time and dataset size could potentially better-align Llama-3.2-3B-Instruct responses with GPT 4.1-mini, effectively allowing you to get GPT 4.1-comparable performance from a 3 billion parameter model.

5.1 Lessons and Issues

Despite the potential implications of fine-tuning a small language model to generate approaches to research questions, it is important to consider whether the NLP task itself is flawed. For one, publishable research often takes novel approaches to research questions, meaning that a large element of creativity is required. Although LLMs can be creative, it is unreasonable to expect them to generate novel research approaches when their training data contains primarily standard research approaches. This shortcoming is present even in today's state-of-the-art models.

Additionally, different research questions can often require vastly different approaches, even if two questions are in the same field. This can make it difficult for an LLM to generalize for such a task. Additionally, given the time and hardware constraints, it is even more unreasonable to expect a small model (around 3 billion parameters) to perform well.

In terms of personal lessons, I've never fine-tuned an LLM previously, so the entire process from dataset creation to model fine-tuning was a learning experience. Additionally, I gained a better understanding for how LoRA works, as well as other parameter efficient fine-tuning methods.