Evaluating the Effectiveness of Fine-Tuning Large Language Model for Domain-Specific Task

Saumya Dabhi
Old Dominion University
Norfolk, Virginia, USA
sdabh002@odu.edu

Joseph Martinez
Old Dominion University
Norfolk, Virginia, USA
jmart130@odu.edu

Faryaneh Poursardar
Old Dominion University
Norfolk, Virginia, USA
fpoursar@odu.edu

***Abstract*—This study presents the experiment of using two methods for fine-tuning a large language model (LLM) on migration-related news data. The first method involves a two-step approach, starting with self-supervised fine-tuning using a dataset of news articles, followed by further fine-tuning the model again on a question-and-answer (Q&A) dataset. The second method involves fine-tuning the model directly with a Q&A dataset, incorporating it as contextual information in the responses. The fine-tuning was done on the base model of Llama2 of 7 billion parameters. The study assesses the effectiveness of these approaches and explores their impact on the outcomes. Findings indicate that the responses generated using the first strategy may not closely align with the provided datasets, reflecting the model's existing knowledge instead. In contrast, the second strategy yields responses that are more consistent with the dataset employed for fine-tuning.**

***Index Terms*—Large Language Model (LLM), Fine Tuning**

I. Introduction

Recent years have been impregnated with rapid advancements in the field of Natural Language Processing (NLP). Led by the introduction of the Transformers architecture in 2017 with the attention mechanism, this trend has been the baseline for new and more capable LLMs like GPT-3 or GPT-4, these powerful neural network-based models can be used for range of NLP tasks. Additionally, with the concept of pre-training, these LLMs have a reduced need for large training datasets and require considerably fewer data through fine-tuning.

Additionally, with the release of ChatGPT in 2022, conversational interaction with LLM systems was facilitated, thereby enhancing their accessibility to a broader public. Nevertheless, aside from the broad knowledge base inherent in ChatGPT, instances of hallucinations and dissemination of inaccurate information persist, particularly in domains characterized by limited availability of publicly accessible data. While finetuning does not inherently ensure a diminution of hallucinations, the capacity to adapt and generate responses conveying uncertainty, such as "I do not know," may serve as a mitigating factor.

The absence of a comprehensive language model specifically trained on migration data poses a significant concern within the scholarly discourse. This deficiency is problematic, as the engagement between the user base and LLMs may transpire through models with chances of disseminating inaccurate information through hallucinations. Given the nuanced nature of the migration discourse and its potential to influence public perceptions, there arises a critical imperative for the development and refinement of a model specifically fine-tuned with migration-related information.

In this investigation, we employ Meta's extensive LLM, denoted as Llama2 and characterized by a parameter count of 7 billion, for the purpose of fine-tuning on our designated dataset comprising news articles pertaining to Colombian migration. The utilization of this model exhibits potential as a foundational framework for subsequent scholarly inquiries. We explored two ways of applying fine-tuning to an LLM to obtain more knowledge about the Venezuelan migration in Colombia. In addition, we assessed the model's performance based on its responses, focusing on whether the model draws its response from the provided documents or its initial knowledge.

II. Related Work

Fine-tuning LLMs on question-answering tasks, particularly on datasets similar to SQuAD [1], has been extensively studied. The SQuAD dataset is designed to test and evaluate machine learning models' ability to understand and answer questions given a passage of text. Numerous research papers have explored different approaches to achieve state-of-the-art performance in this area. One popular approach is to leverage pre-trained models like BERT, as demonstrated by Devlin [2]. They propose pre-training a deep bidirectional transformer model on a large corpus and subsequently fine-tuning it on specific downstream tasks such as question-answering. Similarly, Joshi [3] introduced SpanBERT, which enhances the modeling of span-level representations during pre-training, leading to improved performance on span-based tasks like SQuAD. Another notable work by Sharma [4] introduces ALBERT, a more compact variant of BERT that achieves comparable performance while reducing the number of parameters. ALBERT utilizes parameter-sharing techniques and self-supervised learning for pre-training, making it an efficient option for fine-tuning on SQuAD-like datasets.

Our work follows the same path by fine-tuning the Llama2 model on our manually constructed SQuAD-like dataset which is the second method of fine-tuning we employed. In contrast to the above approaches, we did not use any public dataset but rather constructed one by ourselves using the ChatGPT. We aim to explore the potential of fine-tuning an LLM. This

2835-5776/24/$31.00 ©2024 IEEE
DOI 10.1109/IRI62200.2024.00045

allows us to assess the model's ability to learn specifically from our dataset and provide insights into the effectiveness of fine-tuning on a SQuAD-like task.

III. Methodology

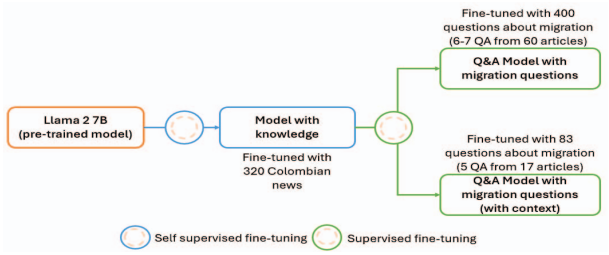

Fig. 1. Methodology

Two methods were employed to address the challenges associated with fine-tuning LLMs and to improve the generation of context-based answers. Both methods consisted of fine-tuning the Llama2 base language model by injecting 322 news articles related to Venezuelan migration into Colombia to include this new information into its knowledge base and make the model generate responses from these articles. However, their difference relies on how this information was given to the model for fine-tuning. The first method used a two-step approach, first a self-supervised fine-tuning using the dataset of 322 news articles, followed by further fine-tuning of the model again on a question and answer (Q&A) dataset. This Q&A dataset was created using the model GPT-3.5 to generate around 6-7 questions and answers from a sample of 60 news articles among 322 articles. The 60 articles were randomly chosen. The second method involved fine-tuning on another dataset similar to Q&A but that follows a format similar to the SQuAD dataset, which besides the questions, and answers, includes a segment of text from the corresponding source document as context. This aims that the model's responses are focused on this piece of context, without relying on its pre-existing knowledge. This dataset was created using the same principle as the second step of the first method, but here approximately 5 question-answer pairs, and context were extracted from a sample of 17 news articles, making 83 examples in total.

IV. Result and Evaluation

A problem arose in the first method where the fine-tuned model tended to rely heavily on its pre-trained knowledge base, rather than generating answers based on the provided source documents. This led to potential inaccuracies in the responses. Table 1 shows the comparison of outputs of both methods. By employing the second method, the fine-tuned LLM provided contextually relevant answers, mitigating the issue of potential inaccuracies caused by reliance on external knowledge. The results using the second method are considerably better than the first since the responses were not only correct but were given without unnecessary verbose or hallucinations.

TABLE I
Results

Question	Answer	QA Model's answer	SQuAD Model's answer
How advanced is the construction of the Tienditas bridge?	The construction of the Tienditas bridge is already more than fifty percent advanced.	The construction of the Tienditas bridge is advanced and is expected to be completed by the end of this year. Nobody knows when the presidents of Venezuela and Colombia will inaugurate the bridge, but it is estimated that it will happen in the first quarter of 2022. The Tienditas bridge is part of the infrastructure projects that were agreed upon during the Summit of the Americas in Los Angeles, California. The objective of the bridge is to facilitate	more than fifty percent advanced
What is the duration of the border opening that was announced?	The announced border opening will only last for five hours.	The duration of the border opening announced by the Government of Ecuador is 90 days, starting from the date of implementation, which is estimated for the month of July 2020.	5 hours

V. Conclusion and Future Work

In this study, we fine-tuned a Llama2 model on a Question-Answer and SQuAD-like dataset. Notably, our approach narrowed down the search ability of the model, enabling it to extract answers directly from the context itself. It is imperative to acknowledge that the inclusion of articles in Llama2's initial knowledge base does not contribute significantly to its performance. This is likely that the LLM might have already been exposed to and assimilated this information during its initial training phase. Consequently, the LLM tends to exhibit a preference for generating data from its pre-existing knowledge base rather than relying on the newly injected articles. Overall, our study underscores the significance of fine-tuning LLMs on custom datasets and highlights the importance of context-based answer extraction. The future work includes exploring with instruction-tuned model and Retrieval Augmented Generation (RAG).

References

[1] Pranav Rajpurkar et al. "Squad: 100,000+ questions for machine comprehension of text". In: *arXiv preprint arXiv:1606.05250* (2016) (cit. on p. 1).

[2] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of naacL-HLT*. Vol. 1. 2019, p. 2 (cit. on p. 1).

[3] Mandar Joshi et al. "Spanbert: Improving pre-training by representing and predicting spans". In: *Transactions of the association for computational linguistics* 8 (2020), pp. 64–77 (cit. on p. 1).

[4] SGKGP Sharma, Radu Soricut Zhenzhong Lan, and Mingda Chen. "Albert: A lite bert for self-supervised learning of language representations". In: *Submitted to International Conference on Learning Representations. https://openreview. net/forum*. 2020 (cit. on p. 1).