AUGMENTING PHI-3 FOR ARABIC: GRANTING LLMS MULTILINGUAL CAPABILITY WITH MINIMAL TRAINING TIME

Omar Abdulbagi
King Abdulaziz University
omarfabdulbagi@gmail.com

Ahmed Alnasser
King Fahd University of Petroleum and Minerals
a3alnasser@gmail.com

Mohammad Alzahrani
University Of Jeddah
moai.ksa1@gmail.com

Bassel Koshak
King Fahd University of Petroleum and Minerals
be@koshak.net

Yasir Alghamdi
King Fahd University of Petroleum and Minerals
Yasir_Alghamdi@outlook.com

ABSTRACT

This project presents an innovative approach to expanding the vocabulary of pre-trained Large Language Models (LLMs) for Arabic without the high costs associated with full retraining. By combining vocabulary expansion, Low-Rank Adaptation (LoRA), and focused fine-tuning, we aim to enhance model performance while minimizing computational demands. This method establishes a new standard for efficient language model improvement in low-resource environments.

1 Introduction

LLMs have shown remarkable capabilities in natural language processing, yet their performance often lags in languages with limited training data, such as Arabic. This project presents an innovative approach to expanding the vocabulary of pre-trained LLMs for Arabic without the high costs associated with full retraining. By combining vocabulary expansion, LoRA, and focused fine-tuning, we aim to enhance model performance while minimizing computational demands. This method addresses key challenges in multilingual AI development, including resource efficiency, linguistic adaptability, and performance enhancement. Our work contributes to the growing research on efficient multilingual model adaptation and establishes a new standard for improving language model performance in resource-constrained environments. By demonstrating the effectiveness of this approach with Arabic, we hope to inspire similar efforts for other underrepresented languages, ultimately working towards more linguistically diverse and capable AI systems.

2 Related Work

2.1 Low-Rank Adaptation (LoRA)

Low-Rank Adaptation (LoRA) Hu et al. [2021], is an efficient fine-tuning technique for large language models that significantly reduces the number of trainable parameters while maintaining model quality. The key innovation of LoRA is to freeze the pre-trained model weights and inject trainable rank decomposition matrices into each layer of the model architecture. By decomposing the weight update matrix into two low-rank matrices, LoRA achieves parameter-efficient fine-tuning, dramatically reducing memory usage and computational costs during training.

The efficiency and minimal computational requirements of LoRA make it an ideal choice for our projects. The implementation of LoRA on the Phi-3 mini model has limited Arabic language capabilities and contains only the basic characters of the Arabic script. This approach focuses on modifying only the embedding layer to incorporate Arabic-specific embeddings while preserving the pre-trained weights and layers of the Phi-3 mini model. This targeted adaptation allows us to leverage the model's existing knowledge and reasoning capabilities while efficiently expanding its understanding of the Arabic language. Using LoRA, we can achieve this adaptation with significantly fewer trainable parameters and reduced computational resources compared to full fine-tuning, making it feasible to enhance Phi-3 mini's Arabic capabilities even with limited computational resources.

2.2 Chinese LLaMA

In their paper, Cui et al. [2023] address the limitations of existing large language models aimed mainly at English audiences in handling non-English languages, specifically Chinese. They introduce an approach to enhance the LLaMA model's ability to process Chinese text by expanding its vocabulary with additional Chinese tokens by training their tokenizer. This extension significantly improves the model's efficiency in encoding and decoding Chinese, thereby enhancing its overall proficiency in understanding and generating text in the new language. Furthermore, they employ the LoRA method, which not only enables efficient fine-tuning without the need for excessive computational resources but also maintains LLaMA's general reasoning abilities and utilizes it to understand Chinese.

In our work, we closely follow Cui et al. [2023]'s methodology to develop an Arabic language model using the Phi-3 mini model. We have adapted their approach by expanding the Arabic vocabulary and applying LoRA for fine-tuning, ensuring that the model is well-equipped to handle the complexities of the Arabic language. This approach has proven instrumental in adapting the LLaMA model's strengths to the Arabic language, ensuring consistency in methodology and enhancing the model's performance across linguistic contexts.

2.3 Phi-3 mini

The Phi-3 family of small language models (SLMs), developed by Microsoft Azure Blog Team [2023], is designed for efficient natural language processing tasks, offering a lightweight architecture that delivers robust performance while minimizing computational demands. Among this family, the Phi-3 mini model stands out due to its compact size and efficiency, making it particularly suitable for applications where resource constraints are a concern. This model's design allows for the development of robust language models without the need for extensive resources, making it an ideal choice for creating an Arabic language model.

We specifically selected the Phi-3 mini model due to its lightweight nature, allowing us to develop a lightweight yet effective Arabic LLM. By implementing the fine-tuning methodology from Cui et al. [2023]'s work on Chinese LLaMA and Alpaca, we adapted the Phi-3 mini to handle Arabic text. This approach leverages the model's lightweight design, ensuring that it remains resource-efficient while providing strong performance in Arabic language processing.

3 Datasets

In this study, we utilized a diverse range of datasets, totaling over 20 gigabytes of Arabic text data Table 1, to ensure comprehensive coverage of the Arabic language and facilitate robust model performance. These datasets contribute to training over 5 billion tokens, which are essential for capturing the linguistic diversity and complexity of Arabic, enabling the model to achieve a higher level of proficiency and adaptability across various tasks.

Because we are implementing LoRA, it was important to use only high-quality data to maximize training efficiency. LoRA allows us to fine-tune large models with fewer resources by adapting only a small number of parameters, making it crucial that the data fed into the model is of the highest quality to achieve optimal performance Hu et al. [2021]. Following the approach of Zhang et al. Cui et al. [2023], we carefully filtered our datasets to include only the most valuable and well-structured Arabic texts. This ensures that the model learns from relevant and representative data, optimizing performance while keeping computational demands low.

3.1 Preprocessing

To maintain consistency across these datasets, a uniform preprocessing protocol was applied. This involved filtering out all non-Arabic characters, including non-Arabic words, and normalizing the text to a standard format. Specifically, we: Removed non-Arabic characters and symbols, ensuring the text strictly adhered to Arabic script. Stripped diacritics were necessary to focus on the fundamental linguistic structure. Split excessively long sentences into manageable parts, preserving their coherence and readability. This preprocessing ensured that all datasets conformed to a consistent format, optimizing them for subsequent stages of tokenization and model training.

4 Training

The training process for our model is designed to build upon the existing capabilities of the base language model while adapting it to handle Arabic effectively. Given the complexity of the Arabic language and the scale of the data, this process involves several key stages, starting with tokenization, followed by pretraining and fine-tuning.

This section details each step of the training process, from tokenization to fine-tuning, and explains how we integrated these methods to create an effective Arabic language model.

4.1 Tokenization

The initial phase of the training process involves the generation of Arabic tokens, which requires expanding the model's vocabulary to include Arabic linguistic elements. This process is fundamental to enabling the model to effectively

Dataset	Tokens	Description
OSCAR	1.7B	OSCAR is a massive multilingual dataset derived from filtered and deduplicated Common Crawl data. It includes a substantial portion of Arabic text, serving as a broad base for language modeling by capturing a wide range of Arabic dialects and styles.
Tahskeela	70M	Tahskeela is a specialized dataset focused on Arabic text with diacritics, providing essential data for training models to handle the intricacies of Arabic orthography. The inclusion of diacritics is crucial for enhancing the model's understanding of word meanings and pronunciations.
ArWIKI	140M	ArWIKI is extracted from the Arabic version of Wikipedia and comprises encyclopedic content that reflects both formal and contemporary Arabic usage. This dataset offers a rich resource of factual and well-structured text, contributing to the model's ability to process and generate coherent Arabic sentences.
UN	100M	The first parallel corpus composed from United Nations documents published by the original data creator. The parallel corpus presented consists of manually translated UN documents from the last 25 years (1990 to 2014) for the six official UN languages, Arabic, Chinese, English, French, Russian, and Spanish. Ziemski et al. [2016]
HARD	300M	The HARD dataset consists of a large collection of Arabic news articles, introducing the model to diverse topics and writing styles present in media, from formal reporting to opinion pieces. This variety is key to enhancing the model's adaptability to different Arabic contexts.
CIDAR	-	CIDAR (Culturally Aligned Instruction Dataset for Arabic) is an instruction-tuning dataset designed to improve language models' performance in Arabic by addressing cultural biases. It includes 10,000 instruction and output pairs, reflecting the diverse cultures of the Arab region. By using CIDAR, our model benefits from instruction-tuning that is better aligned with Arabic grammar and cultural context Alyafeai et al. [2024].

Table 1: Summary of datasets used in the project

understand and process Arabic text. The embedding layer was resized to accommodate the new tokens, ensuring that the model could represent Arabic words with the appropriate level of granularity. To generate these tokens, a variety of sources were used, including extensive Arabic corpora. The tokenization process was carefully designed to capture the unique characteristics of Arabic, such as its script, morphology, and syntax. This step is essential in laying the foundation for the model to perform well in subsequent stages of training.

4.2 Pre-training

After tokenization, the model enters the pertaining phase, where it is pre-trained on a diverse set of Arabic datasets using the LoRA technique. This phase is crucial because it allows the model to build a foundational understanding of the Arabic language while preserving the general reasoning capabilities it developed during its initial training. LoRA is particularly effective in this context as it enables the model to adapt to new languages without requiring full retraining, thus saving time and computational resources. The datasets used during this phase include OSCAR, a massive crawled and filtered dataset that provides a broad representation of Arabic text, and ArWIKI, which is the Arabic version of Wikipedia. These datasets help the model to learn various linguistic patterns, contexts, and nuances specific to Arabic. Additionally, this approach is inspired by similar successful efforts, such as the work by Chinese researchers who used LoRA to efficiently adapt the LLaMA model to Chinese language tasks Cui et al. [2023]. Their success underscores the effectiveness of LoRA in extending language models to new languages with minimal retraining.

4.3 Instruction based fine tuning

Following the pretraining phase, the model is further refined through instruction-based fine-tuning using the CIDAR instruction dataset, which is tailored specifically for Arabic. This step is particularly focused on enhancing the model's ability to follow complex instructions in Arabic, enabling it to perform tasks that require a deep understanding of both the language and the context of the instructions. Fine-tuning is a critical step that adjusts the model's parameters to improve its performance on specific tasks. The CIDAR dataset plays a key role here, as it provides a diverse range of instruction-based tasks that challenge the model's comprehension and execution abilities in Arabic. This fine-tuning

process is meticulously designed to ensure that the model not only understands the instructions but also executes them accurately. The effectiveness of this fine-tuning is rigorously evaluated using the OALLM leaderboard, which serves as a benchmark to assess the model's capabilities. The leaderboard provides a competitive and standardized environment to measure how well the model performs against other language models, offering insights into areas where further improvements might be needed. Through this process, the model's proficiency in instruction-following tasks in Arabic is significantly enhanced, making it a powerful tool for a wide range of applications.

5 Conclusion and Future Work

In conclusion, our work presents a novel approach to augmenting LLMs for the Arabic language, leveraging vocabulary expansion, LoRA, and focused fine-tuning to enhance model performance with minimal computational demands. Despite significant progress in developing and preparing our methodology, we faced challenges in securing adequate computing resources. As a result, we have not yet been able to train the model. However, the code and setup are ready, and once computing resources, such as an NVIDIA A100 GPU, become available, the training process is expected to take approximately three days.

Upon completion of the training, we plan to benchmark the results on the OALLM leaderboard to evaluate the model's performance and identify potential areas for improvement.

For future work, we aim to extend this approach by utilizing a larger model with more parameters. Additionally, expanding the dataset to incorporate more diverse and comprehensive Arabic text sources will likely improve the model's ability to generalize across various tasks and contexts. This continued development will contribute to further advancements in multilingual and low-resource language model capabilities.

References

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv*, 2021. URL https://arxiv.org/abs/2106.09685.

Yiming Cui, Ziqing Yang, and Xin Yao. Efficient and effective text encoding for chinese llama and alpaca. *arXiv*, 2023. URL https://arxiv.org/abs/2304.08177.

Azure Blog Team. Introducing phi-3: Redefining what's possible with slms. https://azure.microsoft.com/en-us/blog/introducing-phi-3-redefining-whats-possible-with-slms/, 2023.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, 2016. European Language Resources Association (ELRA). URL http://conferences.unite.un.org/UNCorpus. Freely available under a liberal license.

Zaid Alyafeai, Khalid Almubarak, et al. Cidar: Culturally aligned instruction dataset for arabic. *arXiv preprint arXiv:2402.03177*, 2024. URL https://arxiv.org/abs/2402.03177.