feedback on the paper : ALLaM: Large Language Models for Arabic and English

Asmaa.Boudjenane

January 2025

The paper outlines several significant contributions of ALLaM:

1 Open-Source Status

ALLaM is an open-source Arabic Large Language Model (LLM) developed by the Saudi Data and Artificial Intelligence Authority (SDAIA). It is hosted on IBM's Watsonx platform, where it is available under a royalty-free SDAIA license, coupled with the Llama 2 Community License. This allows both commercial and government organizations to use and build upon the model without restrictions beyond those in the Llama 2 license .

Accessibility:

Users can access ALLaM via the Watsonx.ai studio, which provides tools for training, fine-tuning, and deploying the model.

2 Approach

Adaptation of Existing Models:

The team first demonstrates the feasibility of adapting a pretrained English model (Llama-2) to handle both Arabic and English. This is done through tokenizer and vocabulary expansion.

Training from Scratch:

After proving the feasibility, the team applies their learnings to train a stronger model from scratch (starting with random initialization). This model is pretrained on English and then further trained on a mix of Arabic and English data.

Tokenizer and Vocabulary Expansion:

The paper highlights the use of tokenizer and vocabulary expansion to adapt an existing English model (Llama-2) to Arabic. This approach minimizes catastrophic forgetting in English while achieving fluency in Arabic .

Model Series:

The ALLaM series includes four models at three different scales: 7B, 13B, and 70B models initialized using Llama-2 weights.

A 7B model trained from scratch (random initialization).

Performance:

The resulting models achieve state-of-the-art results in Arabic and also improve the English performance of the original Llama-2 model.

Training Methodology:

Continued Pretraining: Adapting Llama-2 weights to Arabic and English.

Training from Scratch: Developing a 7B model from random initialization, pretrained on English and fine-tuned on mixed Arabic and English data .

Data Collection: ALLaM was trained on a massive dataset of 3 trillion tokens, including 500 billion Arabic tokens collected from web crawls, books, news articles, and translated English content.

3 ALLaM with other prominent language models

ALLaM : stands out for its state-of-the-art performance in Arabic benchmarks and its focus on cultural alignment for the Arabic-speaking world while also improving English performance over Llama-2.
Jais :is competitive in both Arabic and English but, it is more general-purpose and less culturally aligned.
GPT, Falcon, and BLOOM: have limited Arabic capabilities, with GPT being primarily English-centric.