Abstract

One of the significant problem is the lack of open-source, generative-based small language models available for the Azerbaijan language, an issue which remains despite the existence large-scale, privately funded offerings such as OpenAI GPT-4 or Claude Opus. Given the urgent demand for more serviceable models with the potential to both produce and comprehend Azerbaijani, this study marks the first-ever endeavor to pre-train a generative-based small language model in a targeted manner for this minority language.

While existing large models are proprietary and the trend is moving towards open-weight or open-source Large Language Models with Fine-tuned weights for specific tasks, the best-case scenario is given developed model. Unlike many other languages, it is virtually impossible to fine-tune a dedicated model in Azerbaijani in the absence of either comprehensive or clean datasets for pre-training powerful models to be measured in billions of tokens.

Considering the outlined issues, the research employed the collection of the largest Azerbaijani text corpus available to date, estimated at nearly 3 billion tokens, from a breadth of sources, including Wikipedia, OSCAR, mC4, books, news, and other scraped data. The dataset curation process was thorough, and many filtering steps were taken, including data sampling and preparation, prior to the model training. A 150-million-parameter decoder-only pre-trained language model was then created with the use of Llama2 architecture to facilitate autoregressive text generation in Azerbaijani. Afterward, the model was instruction-fine-tuned with the Alpaca instruction dataset, consequently turning it into a question-answering model in Azerbaijani.

Although the model is less powerful, as evidenced by its smaller size containing only 150 million parameters, compared to larger multi-billion parameter models, it is unique in its ability to generate text in Azerbaijani with few mistakes and answer simple questions. This is a significant outcome for researchers given that the other open-source models, including the largest multi-billion ones with billions of parameters trained on trillions of tokens, cannot demonstrate such linguistic proficiency in the Azerbaijani language. As such, the present study is a vital first step in the field of Azerbaijani language processing, allowing for potential improved performance. The work presents a foundation for further research and applications of fine-tuned generative models in NLP for low-resource languages. It also confirms the possibility of creating effective language models and shows that the size of model training data and computational resources could be easily increased by the scaling laws of machine learning. Thus, the described project fills a notable gap in Azerbaijani language processing and gives a way to research and pro- duce further applications for NLP. The work could benefit the general NLP area by showing the potential of fine-tuned generative models to be used as a tool for enhancing language understanding and multimodal generation and shedding light on prospects for their further improvement and applications for other low-resource languages.

The source code and supplementary materials for this project are available on GitHub at https://github.com/eljanmahammadli/AzLlama for replication and further research.