**Romanian Open-Source Chatbot**

**Candidate: Andrei - Răzvan, JOLDEA**

**Scientific coordinator: Associate Professor Cosmin, CERNĂZANU**

Session: June 2024

**Table of Contents**

**(Arial 20 pt, Bold, Uppercase, Center)**

The abstract is intended to inform about the content of the paper through a brief description of the research of up to one page, the procedures/methods, as well as its results or conclusions. The abstract in Romanian becomes mandatory for works edited in languages other than Romanian and will be written in 12 pt Arial fonts. It will start two blank lines after the heading "ABSTRACT". Before the title, there will be three blank 12 pt. lines.

**ABSTRACT**

**(Arial 20 pt, Bold, Uppercase, Center)**

The English abstract will be on the third page of the manuscript and will synthetically present the paper work. The maximum length of the abstract is one page written with Arial fonts, size 12 pt. The abstract text will begin after two blank lines (size 12pt.) from the “ABSTRACT” title. Before the title there will be left three 12 pt blank lines.

# INTRODUCTION

Over the past few years, the technological landscape has suffered an immense shift due to the emergence of large language models (LLMs). Fuelled by advancements in artificial intelligence (AI) domains such as machine learning (ML), deep learning (DL) and natural language processing (NLP), these models are prevalent in a wide variety of fields, namely healthcare, business intelligence, legal analytics, and even creative arts, encompassing music and literature [1].

While it is true that Large Language Models have impressive capabilities, most of the powerful models are only fluent in the English language, with a few exceptions regarding their multilingual capabilities.

My goal for this thesis was to create a large language model capable of understanding and speaking in Romanian, by finetuning a base model on high quality Romanian data.

A screenshot of a computer

Description automatically generated

Figure 1: GPT-4 benchmarks (source: [2]).

A screenshot of a computer screen

Description automatically generated

Figure 2: Gemini performance on text benchmarks with external comparisons (source: [3]).

A screenshot of a computer

Description automatically generated

Figure 3: Meta Llama 3 Pre-trained model performance (source: [4]).

A screenshot of a computer screen

Description automatically generated

Figure 4: Comparison of Mixtral with Llama (source: [5]).

The Massive Multi-task Language Understanding (MMLU) is one of the most well rounded and popular benchmarks when it comes to large language models. It allows researchers to gain insights into the capabilities and limitations of various LLMs, by measuring the performance of these models using a plethora of tasks and evaluation methods [6].

According to figures 1, 2, 3, 4, we can conclude that the best LLMs on the market right now are OpenAI’s GPT-4 and Google’s Gemini Ultra, each with a 5-shot MMLU benchmark score of 86.4% and 83.7%, respectively. It is worth noting that both these large language models, as well as other highly rated models such as GPT-3.5, PaLM 2L or Claude 2 are closed-source models. On the other hand, Meta’s Llama models and Mistral AI’s Mistral and Mixtral models are open source. This is a relevant aspect that we will discuss further in our analysis. Another detail that is of utmost importance for my task is that these results have been achieved on mainly English tasks.

A graph with numbers and a number of text

Description automatically generated with medium confidence

Figure 5: GPT-4 3-shot accuracy on MMLU across languages (source: [2])

From figure 5, one can infer the fact that the performance of GPT-4 decreases steadily in multilingual scenarios, based on the amount of training data that it had been exposed to during training. This is general amongst all other models.

When choosing between a closed source and an open source LLM, one must take into consideration the following criteria:

* Cost: Closed source models often involve subscriptions in order to have access to the model itself or its API.
* Customization: Open source LLMs are highly customazible for any particular task, due to the fact that the underlying architecture and source code are available to the users.
* Technical expertise: Closed source LLMs are typically more user-friendly, providing an easy to use UI
* Transparency: Users can better understand open source models due to their availability
* Collaboration: Open source communities contribute to the development of the field through shared resources and expertise

Another factor of utmost importance is the computation power available. In this case, I have had at my disposal a virtual machine (VM) with 3 NVIDIA Tesla T4 GPUs, each with approximately 16GB of GDDR6 memory, 16 8-core CPUs, 128GB of RAM and over 2T of storage to complete this task.

As stated before, all the afore mentioned models perform best on English use cases, making a Romanian customization difficult. For instance, if we wanted to create a chatbot specialized in a specific field such as medicine, law or business, a retrieval augmented generation (RAG) using text data from the particular field should be enough. However, since our chatbot should be able to communicate in the Romanian language, this process becomes much more difficult, as it must first be ensured that the LLM is consistent in understaning and providing qualitive answers in Romanian. With this in mind, an open source model would be the most logical choice, since it is highly customizable and would not require additional costs.

A table of numbers and letters

Description automatically generated

Figure 6: Llama 2 language distribution in pretraining data with percentage >= 0.005% (source: [7])

In figure 6, it can be observed the language distribution of the pretraining data for Llama 2. The Romanian language had a 0.03% distribution, which may seem extremely low (and it is) compared to English, for instance, with almost 90%. Even so, my goal could be worse if I was Bulgarian, Danish, Slovenian or Croat, each with a distribution lower than 0.03%.

With all these aspects in mind, it can be concluded that the base model to be used for the finetuning process should respect the following constraints:

* It should be open source.
* It must fit within the available computing resources.
* It must have been exposed to some Romanian data during pretraining

That being said, the large language models that fit these criteria are: Llama-2 7B, Llama-2 13B, Llama-3 8B and Mistral 7B. Altough the official research paper has not yet been released for Llama 3, we can take figure 3 as a reference point and assume that the Romanian language distribution is at least equal to that of Llama 2, but the reasoning capabilities should be much greatear due to its exposure to 15 trillion tokens, compared to Llama 2’s 2 trillion tokens. Therefore, on paper, Llama-3 8B should be the best candidate for this task.

In this paper, I am going to present the techniques, comparisons and results for each of these models, in order to fulfill the task of creating a chatbot that cand follow and express itself in a natural manner in the Romanian language.

**2. Theoretical Foundation**

**Bibliography**

[1] Petroșanu, Dana-Mihaela, Alexandru Pîrjan, and Alexandru Tăbușcă. "Tracing the Influence of Large Language Models across the Most Impactful Scientific Works." Electronics 12.24 (2023): 4957.

[2] <https://openai.com/research/gpt-4>

[3] Team, Gemini, et al. "Gemini: a family of highly capable multimodal models." arXiv preprint arXiv:2312.11805 (2023).

[4] <https://ai.meta.com/blog/meta-llama-3/>

[5] Jiang, Albert Q., et al. "Mixtral of experts." arXiv preprint arXiv:2401.04088 (2024).

[6] Ono, Kensuke, and Akira Morita. "Evaluating Large Language Models: ChatGPT-4, Mistral 8x7B, and Google Gemini Benchmarked Against MMLU." Authorea Preprints (2024).

[7] GenAI, Meta. "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).

[] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[] Lai, Viet Dac, et al. "Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback." arXiv preprint arXiv:2307.16039 (2023).