Polyphemus: Personalized Open-Source Language Models for Yielding Precise Health-Enhancing Medical Understanding and Support

Hayden Spence[1]
[1]Consulting Researcher

Background

Large Language Models (LLMs) are state-of-the-art advancements in natural language processing (NLP) that have revolutionized the field. These models possess remarkable capabilities in comprehending and generating human-like text, pushing the boundaries of language understanding and generation. Recently, the release of the Large Language Model Meta AI (LLaMa) in March 2023 has had a profound impact on the open-source development of LLMs, significantly reducing barriers to entry in this domain. This has led to exciting innovations, such as the Low-Rank Adaptation of Large Language Models (LoRA) project(1), which introduces model updates as low-rank factorizations, effectively reducing the size of update matrices. Another noteworthy project, OpenLLaMA(2), has successfully replicated the functionality of LLaMa under the Apache 2.0 License. These developments, combined with others in the field, have made it feasible to personalize LLMs on consumer hardware.

Methods

Initially, we will establish a data schema to facilitate the collection and augmentation of knowledge from various Observational Health Data Sciences and Informatics (OHDSI) sources, including documentation, tutorials, and educational materials. This process will follow the methodology outlined in the establishment of the RedPajama dataset(3). The collected data will be stored in a document database and used for fine-tuning multiple variants of the OpenLLaMA model, such as the 3B, 7B, and 13B versions. Depending on available computing resources, we may also utilize the LoRA and Parameter Efficient Fine-Tuning (PEFT) methods(4). for the fine-tuning process. To facilitate interactions with the model, existing open-source APIs will be employed to establish a method for utilizing the model in both R and Python terminals.

Second, we will invite the participation of the OHDSI community in a crowd-sourcing effort to contribute real-life healthcare data for reinforcement learning from human feedback (RLHF) to improve the model's performance. This approach will follow the methods described by OpenAssistant(5) and the collected RLHF data will be ingested into the established data schema for further analysis.

Third, we will iteratively update the model based on evaluation results, and to ensure process transparency, these results will be published to Weights & Biases(6). Initially, the focus will be on improving the interpretation of the OMOP CDM v5.4, aiming to generate more accurate SQL data and CDM-specific queries. Once the first specialized model achieves successful results, we will continue this iterative process to develop additional models, incorporating input and feedback from the OHDSI community.

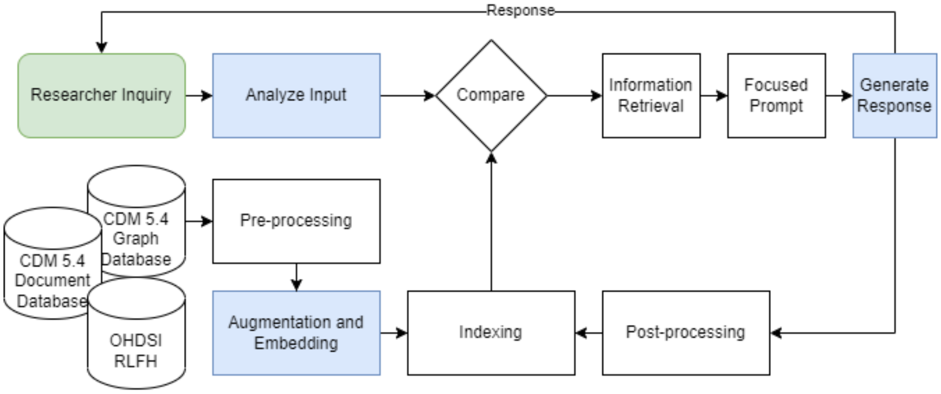

Figure 1. Diagram of the architecture for LLM domain specific information retrieval using the OMOP v5.4 CDM (Green = Human Interface, Blue = LLM Operation)

Lastly, the collection of specialized models will be interconnected, with a general model serving as the entry point. Each specialized model will have access to relevant data sources, environments, and communication methods with peers, enabling them to collaborate and inform the general model's work. This networking approach draws inspiration from the techniques described in the LangChain project[7].

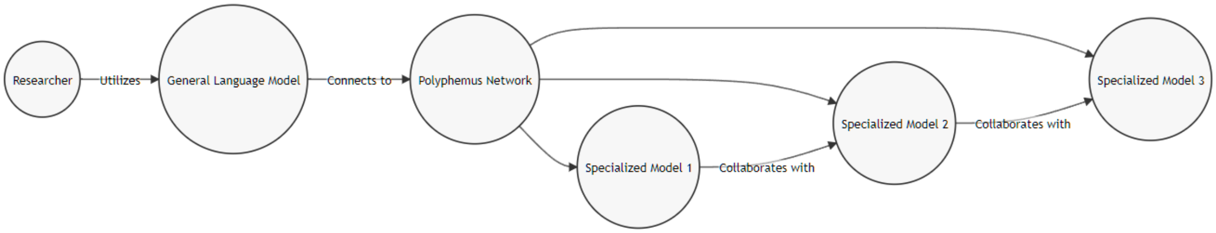

Figure 2. Simplified Diagram of Polyphemus Network

Results

The analysis will focus on three main outcomes. Firstly, we will examine the quantity and types of information ingested by each specialized model. This analysis will provide insights into the effectiveness of the data collection and augmentation process, as well as the relevance of the ingested information to the specific domains addressed by the models.

Secondly, we will assess the level of active involvement from the OHDSI community during the RLHF phase, including their participation in providing narrative user feedback. This evaluation will shed light on the engagement and collaborative efforts within the OHDSI community, which will contribute to the refinement and improvement of the specialized models.

Thirdly, a comprehensive evaluation will be conducted to compare the specialized models against both commercial models and base open-source models. This evaluation will consider both technical aspects and human preferences. It will be performed in both a networked context, where the specialized models collaborate with each other, and an independent context, where the models operate individually. This assessment will provide valuable insights into the performance and capabilities of the specialized models, highlighting their strengths and areas for further enhancement.

Additionally, throughout the development process, a detailed journal will be maintained. This journal will document the challenges encountered and the corresponding solutions devised. This record will serve as a valuable resource for future endeavors and provide guidance to researchers facing similar obstacles.

The ultimate goal is to establish a beta-stage of the Polyphemus framework that requires minimal configuration while delivering substantial value to researchers utilizing LLMs for basic research tasks. Ideally, researchers will be able to leverage a general language model that seamlessly connects to the Polyphemus network of specialized models, enabling them to receive domain-specific responses tailored to their research needs.

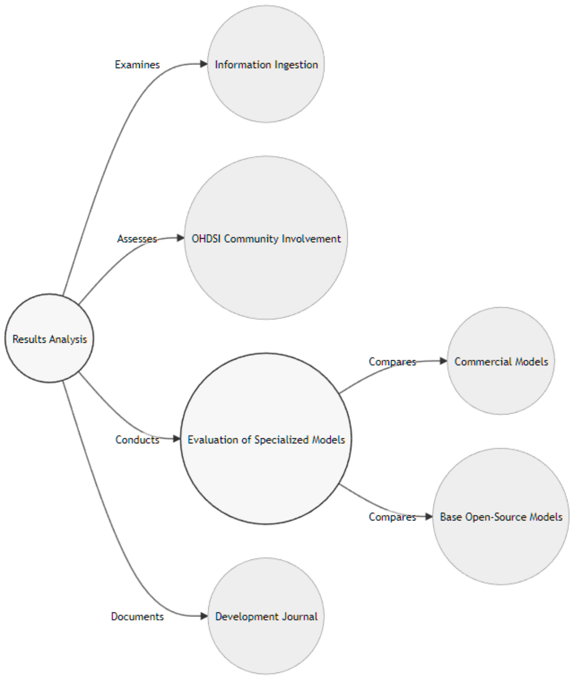

Figure 3. Preliminary evaluation plan

Conclusion

The recent advancements by the research and open source communities following the leaked weights of LLaMa has introduced both democratization and dangers. It is not a question of if major disruptions will take place, it’s a question of how. Without demonstrable use-cases of LLMs being used, it does not matter how much policy is made in academia, government, or industry – people will end up using LLMs because of how effective they are at increasing their productivity. Avoiding the dangers of LLM misuse involves the introduction of alternatives that both perform better in domain specific tasks and have garnered respect as acceptable methods to achieve those tasks.

References

1. Zhang R, Han J, Zhou A, Hu X, Yan S, Lu P, et al. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. ArXiv Prepr ArXiv230316199. 2023;

2. Geng X, Liu H. OpenLLaMA: An Open Reproduction of LLaMA [Internet]. 2023. Available from: https://github.com/openlm-research/open_llama

3. Computer T. RedPajama: An Open Source Recipe to Reproduce LLaMA training dataset [Internet].

2023. Available from: https://github.com/togethercomputer/RedPajama-Data

4. Sourab Mangrulkar SP Sylvain Gugger, Lysandre Debut, Younes Belkada. PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods [Internet]. 2022. Available from: https://github.com/huggingface/peft

5. Köpf A, Kilcher Y, von Rütte D, Anagnostidis S, Tam ZR, Stevens K, et al. OpenAssistant Conversations–Democratizing Large Language Model Alignment. ArXiv Prepr ArXiv230407327. 2023;

6. Weights & Biases – Developer tools for ML [Internet]. [cited 2023 Jun 16]. Available from: https://wandb.ai/site/, http://wandb.ai/site

7. Chase H. LangChain [Internet]. 2022 [cited 2023 Jun 16]. Available from: https://github.com/hwchase17/langchain