Medical Minds Unite: Amplifying Performance by Merging Specialized LLMs

Aditya Chaloo
achaloo@umass.edu

Kedarnath Chimmad
kchimmad@umass.edu

Mayank Bumb
mbumb@umass.edu

Ojas Raundale
oraundale@umass.edu

Siddarth Suresh
siddarthsure@umass.edu

1 Introduction

The field of medicine is witnessing a surge in the application of Large Language Models (LLMs). Trained on vast medical data and literature, these AI systems showcase impressive potential, particularly in facilitating question-answering. LLMs can form the basis of applications, such as virtual medical assistants having the capability of answering complex patient queries, simplifying medical concepts, and providing preliminary diagnoses based on symptoms. While LLMs do possess these capabilities, most of them require to be trained on a huge corpus of data and are proprietary, examples of such models include GPT and BARD. To mitigate these issues with current LLMs, we are proposing two techniques, Model Ensembling and Mixture of Experts. These two methods have proven to improve the performance of weaker open-source models, empowering them to perform at the level of strong proprietary models.

2 Related work

Mixture of experts is a technique that encompasses models that are "experts" in their specific tasks. These experts are activated based on a gating network or router which decides which expert models should receive the tokens. These routes have trainable parameters which are trained at the same time as the rest of the network. We take inspiration from the paper "The Sparsely-Gated Mixture of Experts Layer" ((Shazeer et al., 2017)) which follows the concept of conditional computing where parts of the network will be active based on the input. The Mixture of Expert layer contains a number of feed-forward neural networks with a gating network trained to select a sparse combination of these experts based on the input. Since a sparse combination of the networks is selected, only a few experts are active while the rest are inactive making it computationally efficient. The sparsity is introduced by Noisy Top K Gating where a Gaussian noise is added to the gating network in which top K values are retained while the rest of the values are assigned the value negative infinity resulting in a zero gated value.

Mixtral of experts (Jiang et al., 2024). It is one of the most popular research papers. Mistral.ai has managed to create an LLM with 12B parameters with much better performance than Llama 70B and ChatGPT 3.5 using a concept called Mixture of Experts (Shazeer et al., 2017) (1). The paper takes 8 Mistral (Jiang et al., 2023) decoder models. At each layer, there's a router that passes the inputs to 2 out of the 8 experts which are decided through a gating network. The gating function involves taking the softmax over the Top-K logits of a linear layer. The outputs from the feedforward network of the 2 experts are combined using the weighted sum method. There are 2 mixtral of experts models. One is built using a pretraining model and the other one uses instruction tuning. There's one more concept that is implemented in the paper to save computational resources. The gating vector is designed to be sparse, meaning that for a given input token, most of the weights in the gating vector are zero. This sparsity allows the model to avoid computing the outputs of experts whose corresponding weights in the gating vector are zero, saving computations. We will be using the gating network from this paper to decide the routing for our models.

Biomistral (Labrak et al., 2024). It's an open-source LLM tailored for the biomedical domain that uses Mistral-Instruct as its foundation model and has been further pre-trained on the PubMed Central dataset. It also has been trained on 7 languages marking the first large-scale multilingual evaluation of LLMs in the medical domain.

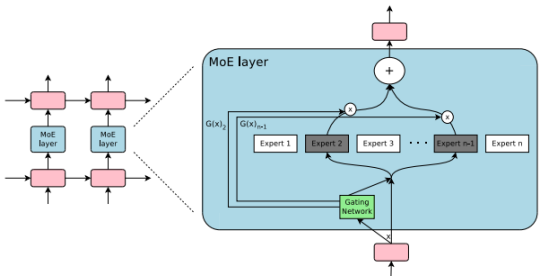

Figure 1: Mixture of Experts (Image from (Shazeer et al., 2017)).

The model architecture of biomistral inherits the standard transformer architecture from Mistral, including features such as Grouped-Query Attention (Ainslie et al., 2023), Sliding Window Attention (Beltagy et al., 2020) and Rolling Buffer Cache. Biomistral has 2 quantized models Activation-aware Weight Quantization (AWQ) (Lin et al., 2023) and BitsandBytes (BnB) which uses QLora (Dettmers et al., 2023) technique. This will help us to get good accuracy with the limited resources that we will be working on. The researchers have also instruction-tuned the model on MCQA (Pal et al., 2022) dataset for evaluation that outperforms a lot of same domain models. We plan to use a quantized version of this model to be one of the experts for the MoE method.

BioGPT (Luo et al., 2022), a domain-specific generative Transformer language model pre-trained on large-scale biomedical literature. BioGPT uses GPT-2 architecture as the model backbone which is a transformer based decoder model. It has been pretrained on the PubMed dataset. And instead of standard GPT-2 vocabulary they learn the vocabulary on the collected in-domain corpus using byte pair encoding (BPE) methodology. And then based on the task it has been finetuned on it.

LLM Blender (Jiang et al., 2023)., a popular LLM Ensembler has architectured a new pairwise ranking system. LLM Blender uses multiple open-source LLMs, ranks them by comparing 2 LLMs against each other, and gives rank by using MaxLogits (a technique that computes the superiority of output compared to other outputs). Out of all the outputs, top K outputs along with the input is fed into a T5 (they used Flan-T5-XL (Chung et al., 2022)) to get a superior output. We are taking the motivation from this paper to fuse the outputs of individual models.

Llama 2 (Touvron et al., 2023), is a collection of pre-trained and fine-tuned large language models (LLMs) optimized for dialogue applications. We are planning to use the Llama 2 7B parameter version. Llama 2 has outperformed most of the LLMs on various different metrics. Their Ghost Attention (GAtt) method has proven to improve multi-turn consistency in dialogues. We believe to capture the general question-answering learning, llama 2 will be helpful and the importance of medical context might be absorbed from the mistral model.

3 Your approach

Our project aims to enhance the performance of biomedical domain-specific tasks through systematic model ensembling. Leveraging three pre-trained models—Biomistral, BioGPT, and Llama 2—we propose to explore various combinations to identify the most effective ensemble. Initial experiments will focus on Biomistral combined with Llama and BioGPT, utilizing a weighted sum method to merge output embeddings. The weights will be adjusted iteratively based on task-specific loss functions, fine-tuning the ensemble until optimal performance is attained. In the second phase, we will introduce a refining layer to further enhance representations before generating the final output, adapting its parameters to task requirements. Subsequently, we will replace this layer with a transformer model in the third iteration, leveraging its self-attention mechanisms to capture intricate relationships within embeddings and refine them accordingly. Each iteration will be evaluated rigorously using standard metrics, aiming to demonstrate significant improvements in classification accuracy and robustness for biomedical tasks. Through this approach, we anticipate not only advancing the state-of-the-art in biomedical classification but also contributing to the broader understanding of model ensembling techniques in specialized domains.

We compare all these accuracies with the original accuracies.

In our next phase, we delve into Mixture of Experts (MoE) techniques to further enhance our model's performance. From the previously mentioned three models—Biomistral, BioGPT, and Llama 2—we will select two for MoE implementation. Post-embedding generation, ensembling will mirror the methods employed in model ensembling. However, MoEs introduce a crucial as-

pect: the routing method. This routing is computed through a gating function, which we will define using a simple layer. This layer will multiply input embeddings with weights, serving as inputs for individual models, and aligning with our chosen ensembling methods.

Additionally, we will explore a second method inspired by the Mixtral of Experts paper. Here, we'll apply softmax over the Top-K logits of a linear layer to design a sparse gating vector. This sparsity optimizes computation by avoiding unnecessary calculations for experts with zero weights, improving efficiency.

We will initiate by pretraining these models on a small subset of the PubMed Central dataset to facilitate better text understanding. Subsequently, fine-tuning will be conducted on a dataset amalgamated from PubMedQA, MedQA, and MedMCQA. This dataset will be partitioned into training and testing sets for evaluation purposes. Through this approach, we aim to harness the collective intelligence of multiple models while optimizing computational efficiency, ultimately advancing performance in biomedical domain-specific tasks.

Our evaluation metric will be accuracy, precision, recall, and f1 score based on how well the model predicts these MCQ-type questions.

One more approach that we plan to try is a question answering for generation tasks wherein the model answers the questions in detail like GPT. We will be using [DATASET] for this task. We will finetune this and use the BART score to evaluate.

3.1 Schedule

We plan to work on everything together. Groups of 2 will tackle 1 model/technique at a time.

1. Preprocessing the data - 1 week
2. Exploring Open Source Models, Ensemble and MoE Techniques - 1 week
3. Building Ensemble Models and MoE - 3 weeks
4. Fine-tuning the model QA Task - 2 weeks
5. Model Analysis and Miscellaneous experiments - 1 week
6. Report Writing - 1 week

4 Data

Pretraining Mixture of Experts layers: For making the Mixture of Experts layers of our proposed model learn world knowledge and understand semantics, contexts and word representations specific to the medical domain, we are planning to use PubMed Central Open Access Dataset. This dataset consists of readily available medical articles and journals textual data.This dataset is published by National Library of Medicine. It consists of CSV files and can be downloaded easily from the FTP server provided here[1].

We have chosen this dataset for pretraining after going through multiple papers on models such as BioMistral and BioGPT, models that we are planning to use for creating our MoE model. These models have been pretrained on the same corpus and it makes sense to us that we pre-train our MoE layers on the same corpus as well.

FineTuning the Ensembled Model and Mixture of Experts Model for detailed question answering:

For further fine tuning or instruction tuning our proposed model on the task of detailed question answering, we are planning to use the PubMedQA[2] dataset (Jin et al., 2019) which is a dataset readily available on Hugging Face. PubMedQA is a novel biomedical question-answering (QA) dataset collected from PubMed abstracts. The task of PubMedQA is to answer research biomedical questions with yes/no/maybe using the corresponding abstracts. PubMedQA has 1k expert-annotated (PQA-L), 61.2k unlabeled (PQA-U) and 211.3k artificially generated QA instances (PQA-A). Each PubMedQA instance is composed of: (1) a question which is either an existing research article title or derived from one, (2) a context which is the corresponding PubMed abstract without its conclusion, (3) a long answer, which is the conclusion of the abstract and, presumably, answers the research question, and (4) a yes/no/maybe answer which summarizes the conclusion.

We are using this dataset for finetuning as it accesses data from the same corpus as PubMed Central Open Access and it makes sense for us to fine tune our models on the task of question answering

[1] https://healthdata.gov/dataset/PubMed-Central-Open-Access-Subset-PMC-OA-/3vwy-a2x4/about_data

[2] https://huggingface.co/datasets/bigbio/pubmed_qa

on a similar corpus for our models to better understand the task of question answering. Further, this dataset has long answers which meets the requirements of our task.

FineTuning the Ensembled Model and Mixture of Experts Model for MCQ question answering:

For further fine-tuning or instruction tuning our model for the task of multiple choice question answering, we are planning to use the MedMCQA[3] (Pal et al., 2022) dataset, which is readily available on huggingface. MedMCQA is a large-scale, Multiple-Choice Question Answering (MCQA) dataset designed to address real-world medical entrance exam questions. MedMCQA has more than 194k high-quality AIIMS and NEET PG entrance exam MCQs covering 2.4k healthcare topics and 21 medical subjects are collected with an average token length of 12.77 and high topical diversity. Each sample contains a question, correct answer(s), and other options that require a deeper language understanding as it tests the 10+ reasoning abilities of a model across a wide range of medical subjects and topics.

We are using this dataset for fine-tuning as it fits our task to make the model learn how to answer multiple-choice questions.

Dataset for Evaluating our models:

For evaluating our models on long question answering, we will be using the Med Quad dataset. The MedQuad dataset[4] provides a comprehensive source of medical questions and answers for natural language processing. With over 43,000 patient inquiries from real-life situations categorized into 31 distinct types of questions, the dataset offers an invaluable opportunity to research correlations between treatments, chronic diseases, medical protocols, and more. Answers provided in this database come not only from doctors but also other healthcare professionals such as nurses and pharmacists, providing a more complete array of responses to help researchers unlock deeper insights within the realm of healthcare

For evaluating our models on MCQ question answering, we will be using the MedAlpaca (Han et al., 2023)[5], medical_meadow_medqa dataset[6].

We are using different datasets for evaluating our model as we want to ensure our models are evaluated on unseen data as far as possible.

5 Tools

The following libraries will be useful for us:

- MergeKit [7]
- BioGPT [8]
- MedAlpacca [9]
- BioMistral [10]
- Lora [11]
- Unsloth.ai (Fast Fine Tune Mistral and Llama) [12]
- Bitsandbytes [13]
- LLM-blender [14]
- LLM Action Aware Weight Quantization [15]

We will be first using our local devices (Laptops with GPUs) and the free version of Google Colab to try out and experiment with the different models. Most probably, we will need further compute for pretraining and fine-tuning all of the above-mentioned models. We will need to purchase Google Colab Pro or Kaggle premium subscriptions for the same.

6 AI Disclosure

- Did you use any AI assistance to complete this proposal? If so, please also specify what AI you used.
  - We used ChatGPT 3.5

If you answered yes to the above question, please complete the following as well:

[3] https://huggingface.co/datasets/openlifescienceai/medmcqa

[4] https://huggingface.co/datasets/keivalya/MedQuad-MedicalQnADataset

[5] https://huggingface.co/medalpaca

[6] https://huggingface.co/datasets/medalpaca/medical_meadow_medqa

[7] https://github.com/arcee-ai/mergekit

[8] https://github.com/microsoft/BioGPT

[9] https://github.com/kbressem/medAlpaca

[10] https://huggingface.co/BioMistral

[11] https://github.com/microsoft/LoRA

[12] https://github.com/unslothai/unsloth

[13] https://github.com/TimDettmers/bitsandbytes

[14] https://github.com/yuchenlin/LLM-Blender

[15] https://github.com/mit-han-lab/llm-awq

- If you used a large language model to assist you, please paste *all* of the prompts that you used below. Add a separate bullet for each prompt, and specify which part of the proposal is associated with which prompt.
  - "Text Containing Our Approach" - Make this text to sound like a project proposal.
  - Please shorten this to 200 words
- **Free response:** For each section or paragraph for which you used assistance, describe your overall experience with the AI. How helpful was it? Did it just directly give you a good output, or did you have to edit it? Was its output ever obviously wrong or irrelevant? Did you use it to generate new text, check your own ideas, or rewrite text?
  - We used AI to change the sound and English structure of our approach to sound more like a project proposal. Later we made changes to it as needed.

References

Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., and Sanghai, S. (2023). GQA: Training generalized multi-query transformer models from multi-head checkpoints.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The Long-Document transformer.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. (2022). Scaling instruction-finetuned language models.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs.

Han, T., Adams, L. C., Papaioannou, J.-M., Grundmann, P., Oberhauser, T., Löser, A., Truhn, D., and Bressem, K. K. (2023). MedAlpaca – an open-source collection of medical conversational AI models and training data.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2024). Mixtral of experts.

Jiang, D., Ren, X., and Lin, B. Y. (2023). LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., and Lu, X. (2019). PubMedQA: A dataset for biomedical research question answering.

Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., and Dufour, R. (2024). BioMistral: A collection of open-source pretrained large language models for medical domains.

Lin, J., Tang, J., Tang, H., Yang, S., Dang, X., Gan, C., and Han, S. (2023). AWQ: Activation-aware weight quantization for LLM compression and acceleration.

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., and Liu, T.-Y. (2022). BioGPT: Generative pre-trained transformer for biomedical text generation and mining.

Pal, A., Umapathi, L. K., and Sankarasubbu, M. (2022). MedMCQA : A large-scale multi-subject multi-choice dataset for medical domain question answering.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. (2017). Outrageously large neural networks: The Sparsely-Gated mixture-of-experts layer.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models.