Merge techniques for large language models

Martynas Baltusis (s222858), Yayi Wang (s243559), Alexander Pittas (s230265), Patrik Suchopa (s241943)

Technical University of Denmark

DTU Compute
Department of Applied Mathematics and Computer Science

Introduction

Large language models (LLMs) have revolutionized natural language processing, but face significant challenges due to their enormous size, high computational demands, and limited accessibility. Fine-tuning smaller pre-trained models and merging them offers an alternative, enabling strong performance with reduced resource requirements.

Our project explores this approach in the biomedical domain by merging models that focus on the biomedical sector and are based on Mistral-7B [1], pre-trained for biomedical tasks. Using methods like DARE (Drop And Rescale), SLERP (Spherical Linear Interpolation), and TIES (Task-Interpolated Embedding Space), we aimed to create adaptable, resource-efficient models. We used the Mergekit [2] tool for merging and different datasets for evaluation, and we seeked to identify the best performing merging strategy for specialized tasks, particularly in healthcare.

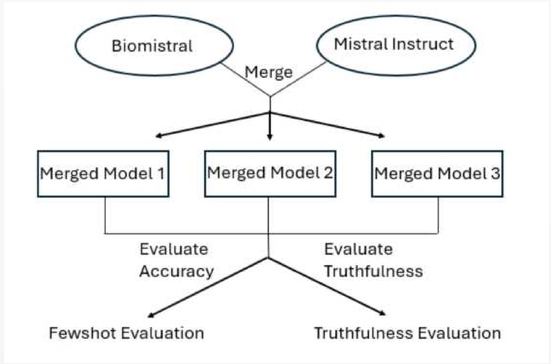

FIGURE 1: Procedure overview

Merging methods

DARE [3] Drop-and-Rescale method assumes a fine-tuned model with parameters θ_{SFT} and a pre-trained model with parameters θ_{PRE} and $\delta^t = \theta_{\text{SFT}} - \theta_{\text{PRE}}$. It has been shown that these δ^t parameters are often highly redundant. The method consists of the following steps:

$$
\begin{aligned}
m^t &\sim \text{Bernoulli}(p), \\
\tilde{\delta}^t &= (1 - m^t) \odot \delta^t, \\
\hat{\delta}^t &= \tilde{\delta}^t / (1 - p),
\end{aligned}
$$

Finally, combining $\theta^t_{\text{DARE}} = \hat{\delta}^t + \theta^t_{\text{PRE}}$ to obtain the parameters for inference. DARE can preserve the model performance by approximating the original embeddings.

TIES [4] Task-Interpolated Embedding Space reflects the fact that features may interfere across models, which results in reduced model performance. This can be a consequence of redundant parameters, parameter "sign disagreement", or both. The method preserves influential parameters by resetting redundant parameters back to the pre-trained values or 0. It then proceeds with a sign election and ends up averaging the values that sign-wise align with the elected sign.

SLERP [5] Spherical Linear Interpolation is a method based on linear interpolation (LERP) that computes intermediate points along the shortest path on a euclidean line, whereas SLERP computes the points along the sphere. It takes two normalized vectors on input, calculates the angle between them and then performs the interpolation. For a pair of normalized vectors $v, w \in \mathbb{R}^n$ and real parameter $\alpha \in [0; 1]$ determining the scale of linear combination between v, w, can be described as follows:

$$
\text{SLERP(v,w,}\alpha) = \frac{\sin((1-\alpha)\theta)}{\sin(\theta)} v + \frac{\sin(\alpha\theta)}{\sin(\theta)} w
$$

where the angle θ is obtained as:

$$
\theta = \cos^{-1}(v \cdot w).
$$

Configuration For our experiments, we have used the tool `MergeKit` [2] with a density value $p = 0.5$ for DARE. DARE, TIES, and DARE-TIES merges had the weight value set to 0.5 per model. For SLERP, the real parameter is set to $\alpha = 0.5$.

Evaluation Accuracy

We evaluated the accuracy of two base models and three merged models using few-shot learning on 10 benchmark datasets.

Models:

Category	Models
Base Models	BioMistral, Mistral-Instruct
Merged Models	BioMistral-Instruct-SLERP, BioMistral-Instruct-TIES, BioMistral-Instruct-DARE-TIES

TABLE 1: Models for evaluation.

BioInstructQA Datasets:[6]

- 4 main datasets: MMLU, MedQA, MedMCQA, PubMedQA.
- 10 evaluation corpora: 6 subjects relevant to medical a clinical knowledge of MMLU, college biology, college medicine, anatomy, professional medicine, medical genetics, and clinical knowledge. MedQA also contains two parts, four-choice (MedQA) and five-choice (MedQA 5 options).

Evaluation Process

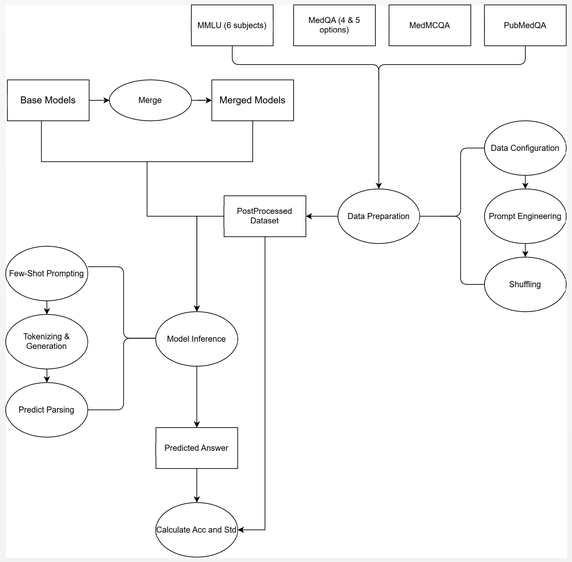

FIGURE 2: Flowchart of few-shot evaluation.

Results

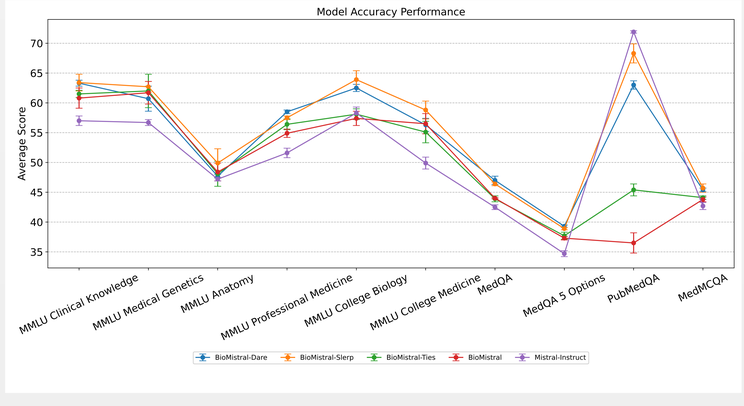

FIGURE 3: Performance of 3-shot in-context learning. The scores represent accuracy.

Truthfulness evaluation & additional model merging

Truthfulness evaluations are performed on LLMs to ensure they do not fascilitate misconceptions and misinformation in their answers. A benchmark for model truthfulness is the TruthfulQA [7] dataset. Focusing on health and medicine related categories (nutrition, health, psychology, science), we performed a single-shot truthfulness evaluation in the form of providing two multiple choice prompts, of which one explicitly prompts the model to answer truthfully, positively, and without bias. In addition, InternistAI [8], a model trained on physician curated data, was added as an optional merge option.

Category	Models
Base Models	BioMistral, Mistral-Instruct-v0.1/-v0.2, InternistAI
Merged Models	BioMistral-Instruct-SLERP, BioMistral-Instruct-TIES
	BioMistral-InternistAI-SLERP, BioMistral-Instruct-DARE-TIES
	BioMistral-InternistAI-TIES, BioMistral-InternistAI-DARE-TIES

TABLE 2: Models for TruthfulQA dataset evaluation.

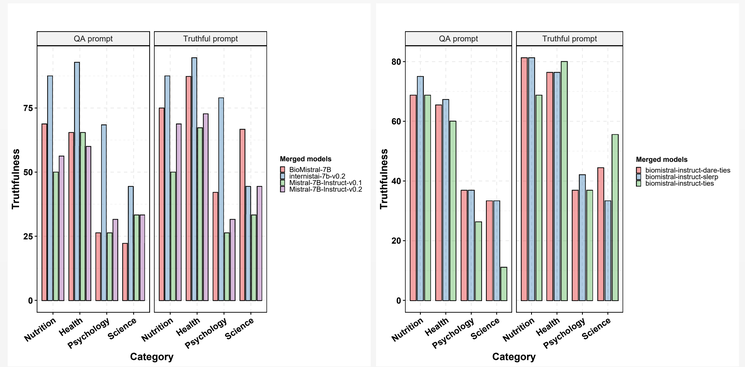

FIGURE 4: Truthfulness comparison for different un-merged LLMs (Mistral-7B-Instruct-v0.1/-v0.2), BioMistral, and InternistAI in health related categories (left) and different merge techniques (DARE/DARE-TIES/SLERP) for Mistral-7B-Instruct-v0.1 / BioMistral (right).

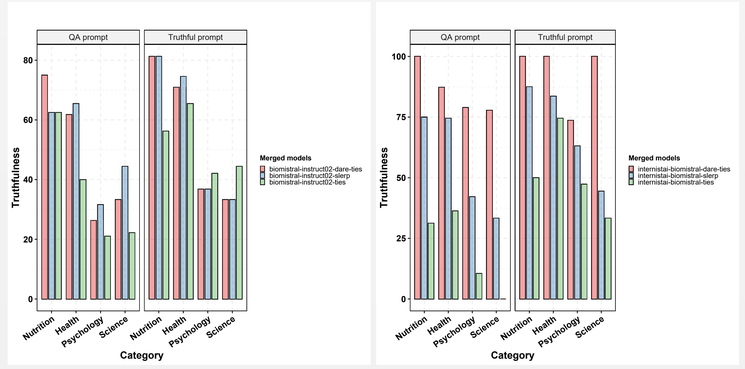

FIGURE 5: Truthfulness comparison for different merge techniques (DARE/DARE-TIES/SLERP) for Mistral-7B-Instruct-v0.2 / BioMistral (left) and BioMistral / InternistAI (right) in health related categories.

References

[1] Devendra Singh Chaplot. Mistral 7b. 2023. URL https://arxiv.org/abs/2310.06825.

[2] Goddard et al. Arcee's mergekit: A toolkit for merging large language models. pages 477–485, 2024. URL https://aclanthology.org/2024.emnlp-industry.36.

[3] Yu et al. Language models are super mario: Absorbing abilities from homologous models as a free lunch, 2024. URL https://arxiv.org/abs/2311.03099.

[4] Yadav et al. Ties-merging: Resolving interference when merging models, 2023. URL https://arxiv.org/abs/2306.01708.

[5] Jang et al. Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval, 2024. URL https://arxiv.org/abs/2405.00571.

[6] Labrak et al. Biomistral: A collection of open-source pretrained large language models for medical domains, 2024. URL https://arxiv.org/abs/2402.10373.

[7] Lin et al. Truthfulqa: Measuring how models mimic human falsehoods. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:3214–3252, 2022. ISBN 9781955917216. ISSN 0736587X. URL https://aclanthology.org/2022.acl-long.229.

[8] Griot et al. Impact of high-quality, mixed-domain data on the performance of medical language models. *Journal of the American Medical Informatics Association*, 31:1875–1883, 9 2024. ISSN 1527974X. URL https://dx.doi.org/10.1093/jamia/ocae120.