# Simple Llama Merge: What Kind of LLM Do We Need?

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Model merging involves integrating multiple specialized models into a single, more powerful model. This approach provides several advantages, including decreased storage and serving costs, enhanced generalization capabilities, and facilitation of decentralized model development. The question of how to effectively combine specialized fine-tuned small models (8B parameters) to achieve performance levels comparable to those of larger models remains an unresolved issue. Therefore, this paper describes our method for the simple merging of models from the LLaMA family. The resulting model is capable of producing complete, instruction-compliant, and highly accurate answers to questions across multiple domains. It achieved 2nd place on the final test of the LLM Merging Competition. For detailed implementation, please refer to our GitHub repository at: `https://github.com/Catrin-baze/llama-merging`

## Introduction

The advent of large language models (LLMs) has transformed natural language processing, yet their effective utilization is often hindered by resource demands and task specificity. Model merging, as defined by Raffel et al.[1], offers a compelling solution by integrating multiple specialized models into a single, more powerful entity. This approach not only reduces storage and serving costs through model reuse across tasks but also enhances generalization by compositional integration of expert capabilities. Additionally, model merging encourages decentralized development, allowing diverse contributors to build and integrate models.

Extensive research in model merging has led to several mainstream methods. SLERP Merging[2] employs spherical linear interpolation to blend expert model parameters smoothly, maintaining continuity and stability during the merging process, especially when models share similar architectures. Task Arithmetic[3] introduces "task vectors," quantifying the differences between fine-tuned expert parameters and base model parameters, enabling a controlled linear combination of expert knowledge that retains essential task-specific information. TIES Merging[4] addresses the challenges of noise accumulation during fine-tuning and parameter conflicts among experts by refining model parameters systematically. Finally, Dare Merging[5] builds on TIES, employing a dropout-like pruning mechanism to reduce noise and focus on significant contributions from each expert model.

Recently, various libraries have emerged to facilitate automated model merging, significantly streamlining the process. One such library is MergeKit[6], which integrates many of the aforementioned methods. Numerous top-ranked models in the Hugging Face Open LLM Leaderboard have been developed using these automated tools. In this study, we utilize MergeKit to develop a model employing a straightforward yet effective spherical linear interpolation (slerp) merging method, complemented by chain-of-thought (CoT) and few-shot techniques. This enables the merged model[1] to generate

[1] `https://huggingface.co/catrinbaze/llama-refueled-merge`

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

comprehensive, instruction-compliant, and highly accurate responses to inquiries across diverse domains.

# Methods

## .1 Model Merging Method

We conducted a comprehensive screening of eligible large language models (LLMs) by initially assessing the performance of each unmerged model on the officially released test datasets. As a result of this evaluation, we selected refuelai/Llama-3-Refueled[2] as one of the parent models due to its superior performance.

Llama-3-Refueled, which is based on the Llama3-8B architecture, effectively integrates over 2,500 unique datasets, thereby ensuring high-quality outputs. This model demonstrates exceptional capability in rapidly adapting to specific domains, exhibiting significant enhancements in label accuracy with minimal training time.

The second parent model selected for merging is meta-Llama/Meta-Llama-3-8B-Instruct[3]. We employed the slerp (spherical linear interpolation) technique to merge these two models, utilizing specific parameter settings: the self-attention layer values were set to [0, 0.5, 0.3, 0.7, 1], and the MLP (multi-layer perceptron) layer values were configured as [1, 0.5, 0.7, 0.3, 0].

## .2 Prompt Engineering

In our investigation, we observed that the prompts provided to the model significantly influence the quality of its generated outputs. To enhance the model's performance, we implemented Chain-of-Thought (COT) reasoning and three-shot learning techniques. Specifically, we employed tailored prompt templates, as illustrated in Table 1, to address various problem types effectively.

# Performance Study

In this section, we delve into several explorations conducted during the competition, addressing the following questions:

Q1: What criteria can be utilized to determine whether a large language model (LLM) achieves superior performance, and which evaluation metrics are employed to measure this?

Q2: What factors influence the scores observed on the contest test set, and to what extent do these factors exert their influence?

## .1 Selection Of Benchmarks

To efficiently evaluate our LLM given limited submission attempts, we referred to evaluation metrics from the OpenLLM Leaderboard[4], such as ARC[7], HellaSwag[8], MMLU[9], TruthfulQA[10], and GSM8K[11]. However, we found that results on these benchmarks do not correlate linearly with performance on the competition test set.

Table 2 shows the benchmark and competition test results for several screened LLMs, including our own merged model, evaluated under Chain of Thought (COT) and 3-shot settings. While models like Hermes-2-Pro-Llama-3-8B[5] and Daredevil-8B[6] performed competitively on standard metrics, they underperformed on the racetest when using our prompt approach compared to the Refueled series[7].

Upon analysis, Hermes-2-Pro-Llama-3-8B and Daredevil-8B tended to produce repetitive statements with our prompt template, possibly due to the sampling parameter being set to FALSE, which restricts

[2] `https://huggingface.co/refuelai/Llama-3-Refueled`
[3] `https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct`
[4] `https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard`
[5] `https://huggingface.co/NousResearch/Hermes-2-Pro-Llama-3-8B`
[6] `https://huggingface.co/mlabonne/Daredevil-8B`
[7] `https://huggingface.co/refuelai/Llama-3-Refueled`

Table 1: Prompt templates used separately for different tasks

| | |
|---|---|
| **Multiple Choice Prompt** | The following are multiple choice questions. Think step by step and then output the answer in the format of "The answer is (X)" at the end.<br>**Question:** The symmetric group $S_n$ has $factorialn$ elements, hence it is not true that $S_{10}$ has 10 elements. Find the characteristic of the ring 2Z.<br>**Options:** A. 0 B. 30 C. 3 D. 10 E. 12 F. 50 G. 2 H. 100 I. 20 J. 5<br>**Answer:** Let's think step by step. A characteristic of a ring is R is $n$ if the statement $ka = 0$ for all $ain2Z$ implies that $k$ is a multiple of $n$. Assume that $ka = 0$ for all $ain2Z$ for some $k$. In particular $2k = 0$. Hence $k = 0$ and $n = 0$. The answer is (A).<br>**Question:** Which of the following is the body cavity that contains the pituitary gland?<br>**Options:** A. Ventral B. Dorsal C. Buccal D. Thoracic E. Pericardial F. Abdominal G. Spinal H. Pelvic I. Pleural J. Cranial<br>**Answer:** Let's think step by step. We refer to Wikipedia articles on anatomy for help. Let's solve this problem step by step. The pituitary gland is the major endocrine gland attached to the base of the brain, and it is contained in the Cranial cavity. The answer is (J).<br>**Question:** Say the pupil of your eye has a diameter of 5 mm and you have a telescope with an aperture of 50 cm. How much more light can the telescope gather than your eye?<br>**Options:** A. 1000 times more B. 50 times more C. 5000 times more D. 500 times more E. 10000 times more F. 20000 times more G. 2000 times more H. 100 times more I. 10 times more<br>**Answer:** Let's think step by step. The amount of light is proportional to the aperture area $A = piD^2/4$ for a lens with diameter $D$, so the relative amounts of light between the eye with diameter 5mm and the telescope with diameter 50mm is $(50cm)^2/(5mm)^2 = 10000$. The answer is (E). |
| **Generation Prompt** | In this task you are given a question. You need to generate an answer to the question.<br>Input: Who was the man behind The Chipmunks?<br>Output: David Seville.<br>Input: On 2 November 2010, the oil painting "Nude Sitting on a Divan" sold for $68.9 million, a record for an artwork by which artist?<br>Output:Amedeo Modigliani |

output diversity. Additionally, Daredevil-8B occasionally generated garbled outputs, while Storm-7B produced extraneous text, such as "Repeat of Request (exact wording)," when prompted to reiterate specific phrases.

These findings suggest that high benchmark scores may not guarantee optimal performance on competition-specific tasks, as models often need specific prompt adjustments for competitive scenarios. While models that excel on the competition test set align well with competitive goals, the effectiveness of the test set as a comprehensive measure of model merging success requires further evaluation. Indeed, our merged model, despite underperforming on some benchmarks, achieved strong results on the test set.

We hypothesize that robust base models like Llama-3-Refueled and Llama-3-8B-Instruct perform well with prompt applications post-merging, while more powerful models may require additional tuning to fully activate their generative potential in response to specific prompts.

Table 2: Performance of LLMs Across Multiple Benchmark Metrics

| Model | ARC | HellaSwag | MMLU | TruthfulQA | GSM8K | Competition Test |
|---|---|---|---|---|---|---|
| Hermes-2-Pro-Llama-3-8B | **0.635** | 0.832 | 0.648 | 0.566 | 0.679 | 0.32 |
| Daredevil-8B | 0.687 | **0.845** | **0.692** | **0.598** | **0.735** | 0.47 |
| Llama-3-Refueled | 0.547 | 0.791 | 0.647 | 0.418 | 0.616 | 0.58 |
| llama-refueled-merge | 0.559 | 0.815 | 0.662 | 0.499 | 0.629 | **0.67** |

### .2 Factors Affecting Results

Table 3 presents the competition test scores for the Llama-3-Refueled model following various experimental treatments. Here, the Refueled-Hermes-2-Pro-Slerp[8] model represents a spherical linear interpolation (slerp) merge of Llama-3-Refueled and Hermes-2-Pro-Llama-3-8B.

The results suggest that merging with certain models may lead to performance degradation, indicating that compatibility between merged models is essential. In contrast, prompt engineering techniques, particularly the use of Chain of Thought (CoT) and few-shot prompting, substantially enhance performance. These findings emphasize the impact of tailored prompting strategies on model effectiveness in competition settings.

The results of our experiments underscore the importance of compatibility among merged models, suggesting that not all combinations yield improved performance. Additionally, they highlight that effective prompt engineering can lead to substantial performance gains, making it a critical area for further exploration. By refining prompts and adjusting the interaction dynamics, we can better leverage the underlying capabilities of each model, resulting in a more effective solution for competition-specific tasks.

In summary, these findings affirm that while high benchmark scores can indicate a model's potential, they do not necessarily correlate with its performance in specialized applications. A holistic approach that includes both model selection and prompt optimization is essential for achieving superior results in competitive environments. Future work should focus on exploring additional prompt variations and assessing their impact on different model architectures to refine our understanding of how best to harness the capabilities of large language models.

Table 3: Test Scores of Llama-3-Refueled Model After Different Treatments

| Model Variant | Competition Test Score |
|---|---|
| Refueled-Hermes-2-Pro-Slerp | 0.47 |
| Llama-3-Refueled (without prompt) | 0.58 |
| Llama-3-Refueled (1-shot) | 0.62 |
| Llama-Refueled-Merge (COT + 3-shot) | **0.67** |

## Conclusions

In this study, we explored the potential of model merging, specifically using spherical linear interpolation (slerp) merging techniques, to enhance large language model performance across diverse tasks. By leveraging MergeKit along with chain-of-thought (CoT) and few-shot prompting methods, we demonstrated that merged models can achieve significant improvements in accuracy, coherence, and instruction compliance. This approach not only optimizes resource utilization but also broadens the applicability of LLMs to more complex, multifaceted tasks. Our findings highlight model merging as a promising avenue for future advancements in LLM development, especially for applications requiring specialized expertise and adaptability.

## References


[1] Colin Raffel. A call to build models like we build open-source software, 2021.


[8] `https://huggingface.co/catrinbaze/refueled-hermes-2-pro-slerp`

[2] Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pages 245–254, 1985.

[3] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

[4] Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Resolving interference when merging models. *arXiv preprint arXiv:2306.01708*, 1, 2023.

[5] Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*, 2024.

[6] Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. Arcee's mergekit: A toolkit for merging large language models. *arXiv preprint arXiv:2403.13257*, 2024.

[7] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457, 2018.

[8] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics*, 2019.

[9] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300, 2020.

[10] Stephanie C. Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Annual Meeting of the Association for Computational Linguistics*, 2021.

[11] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168, 2021.