Is heterogeneous model soup tasty? A Multidimensional Evaluation of Diverse Model Soups in Language Model Alignment

Dawid Motyka[1 [0009−0009−6222−8557]], Paweł Walkowiak[1 [0009−0008−0381−9202]], Julia Moska[1 [0009−0003−8581−1098]], Bartosz Żuk[2 [0009−0008−8473−7718]], Karolina Seweryn[3 [0000−0003−0617−7301]], and Arkadiusz Janz[1 [0000−0002−9203−5520]]

[1] Wrocław University of Science and Technology
{arkadiusz.janz|dawid.motyka}@pwr.edu.pl
[2] Institute of Computer Science, Polish Academy of Sciences
[3] NASK - National Research Institute

Abstract. Training and fine-tuning language models is becoming increasingly expensive. "Model soups" offer a promising solution by combining parameters from separately trained models to create a new one with merged capabilities. Our paper explores using heterogeneous model soups to improve LLM alignment by combining models trained with different alignment methods - a novel approach not previously explored in literature. Through empirical evaluation using an "LLM-as-a-judge" approach, we found that mixing different types of models can improve alignment performance, though this requires careful adaptation of interpolation techniques to account for varying alignment objectives. We've shared our model merging source code on GitHub [4].

Keywords: Model Soups · Alignment · Language Models.

1 Introduction

In recent years, the dynamic development of language models has significantly improved task-specific performance across a wide range of natural language processing tasks. However, the increasing costs of training and fine-tuning large language models for specific applications pose a significant computational challenge for researchers and practitioners. One promising solution to this problem is a model merging technique known as model soups. Model soups combine the parameters of multiple independently trained models to create a new model with merged capabilities and characteristics. This approach offers several advantages, including reduced computational costs, improved adaptation stability, and the ability to personalize models without the need for retraining. As individual models can often be limited by local minima encountered during optimization, combining multiple models should average their capabilities and mitigate these

[4] https://github.com/dawidm/iccs-2025-model-soups

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_19

limitations. This provides a compelling reason for training models multiple times under varying conditions.

Model soups are usually considered in the context of homogeneous models trained with equivalent fine-tuning methods. In contrast, our study explores *heterogeneous* model soups created using diverse alignment methods. We empirically evaluate resultant models formed via DPO, KTO, and ORPO methods, using both linear (LERP) and spherical interpolation (SLERP). Our evaluation focuses on key alignment goals, safety, factual accuracy, linguistic correctness, conciseness, and proactivity. To assess models performance, we established an evaluation framework grounded in a strong language model serving as a judge.
1. We examine the compatibility of DPO-, KTO-, and ORPO-aligned models mixed via LERP and SLERP, assessing their performance across key alignment targets. Heterogeneous mixtures remain underexplored in prior work.
2. Our evaluation utilises a multidimensional framework focused on core alignment goals – safety, factuality, ling. correctness, conciseness, and proactivity.
3. We show, that model mixing techniques such as SLERP, require careful adaptation when combining aligned models.

2 Related Work

Averaging of models' parameters, also called model soups [20], is a widely used approach that showed a range of applications. [7] showed that simply averaging model parameters across its learning trajectory leads to better generalisation. The important concept of *linear mode connectivity* was introduced by [4] who demonstrated that, given shared initialization, two networks tend to converge to the points connected by a line with a relatively flat error rate. [13] showed that the fine-tuned models are similar in the feature and parameter space. [19] and [17] used averaging for an improvement in the out-of-distribution performance. The concept of linear mode connectivity also extends to multiple tasks [12]. [6] demonstrated that *task vectors* can be merged to build a multitask model. [9] also explored the merging of *expert* language models to achieve compositional capabilities.

The less explored field is merging models with different training objectives, which was also shown to be promising. [6] explored merging *task vectors* from both supervised and unsupervised objectives, and [2] used model soups with various loss functions to improve model adversarial robustness.

The concept of model soups has also been employed in the context of the alignment task. [16] investigated RL fine-tuning with interpolation of diverse reward models, while [8] demonstrates the potential to achieve personalized alignment through weighted interpolation of diverse models.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_19

3 Methods

To align the models with human preference data, we used offline alignment, as it allows for direct preference optimization without an explicit reward function, requiring fewer computational resources compared to online methods.

Direct Preference Optimization (DPO) [15] learns implicit reward function by increasing the relative log probability of preferred to non-preferred response, with KL-divergence penalty regularization (reference model).

Odds Ratio Preference Optimization (ORPO) [5] contrastively to DPO, does not incorporate a reference model. It combines the odds ratio between the chosen and rejected responses and a supervised fine-tuning component in its optimization objective.

Kahneman-Tversky Optimization (KTO) [3] – authors proposed an alternative training objective based on Kahneman-Tversky model of human utility, with no need for paired preference data, showing that it exceeds performance of DPO and ORPO in many cases.

In our study, we conducted model alignment using Huggingface TRL [5] implementation of described methods. We used `PLLuM-12B-instruct`[6] base model (further called SFT model) – a 12B parameter model for Polish language, based on `Mistral-NeMo`[7]. All models were trained with AdamW optimizer, learning rate of 2e-6, weight decay of 1e-3 and effective batch size of 64. Regarding method-specific parameters, we used $\beta = 0.1$ for DPO and KTO. $\lambda = 0.2$ for ORPO and $\lambda_D = \lambda_U = 1$ for KTO.

3.1 Model merging

We utilize two popular model merging techniques—Linear Interpolation (LERP) and Spherical Linear Interpolation (SLERP) to combine models trained with diverse alignment methods. We call both approaches model soups. Merges are conducted in whole model parameter space: MLP, attention layers' transformations, and RMS normalizations (RMSNorm) parameters.

Linear Interpolation (LERP) is a method where the weights of two models are combined linearly, specifically by using linear interpolation between weights of 2 models: $\theta_{LERP} = \lambda \cdot \theta_{m_1} + (1-\lambda) \cdot \theta_{m_2}$, where $\theta_{m_1}, \theta_{m_2}$ are parameters vectors of aligned models and λ is the weighting parameter.

Spherical Linear Interpolation (SLERP) is an alternative interpolation method, which preserves norms of parameters vectors [18]. As the parameters of the aligned models are close to the SFT model, we operate on *alignment vectors*: $\theta_{d_i} = \theta_{m_i} - \theta_{SFT}$. Specifically we calculate:

$$\theta_{SLERP} = \theta_{\text{SFT}} + \frac{\sin[(1-\lambda)\Omega]}{\sin \Omega} \cdot \theta_{d_1} + \frac{\sin[\lambda\Omega]}{\sin \Omega} \cdot \theta_{d_2}$$

[5] `https://github.com/huggingface/trl/tree/v0.13.0`
[6] `https://huggingface.co/CYFRAGOVPL/PLLuM-12B-instruct`
[7] `https://mistral.ai/news/mistral-nemo`

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_19

for corresponding parameter matrices with λ controlling influence of source models. To calculate the angle Ω, two-dimensional matrices are reshaped as vectors.

3.2 Model Evaluation

A high-quality evaluation of aligned large language models is both challenging and time-consuming. We use the "LLM-as-a-Judge" approach [21], which can be used to approximate the evaluation of human preferences in a cheaper and faster way. We choose *pairwise comparison* of the human-written gold answer and evaluate the model response. For the judge model, we use stronger LLM – (`Llama3.1-70B`[8]).

We employ the **win-tie-rate (WTR)** metric – the percentage of test cases x in which the response z_t from the evaluated model t is either superior to or on par with the corresponding gold-standard answer z_g, with respect to predefined evaluation criteria. The WTR score for a given response evaluation function Q is calculated as follows: $WTR(T, G) = E_x[\mathbb{1}_{Q(z_t|x)>=Q(z_g|x)}]$, where z_t is the response generated by the evaluated model t, T is the set of model responses $z_t \in T$, G is a set of corresponding gold-standard responses $z_g \in G$.

Evaluation Criteria We selected seven evaluation dimensions that represent typical alignment objectives: safety, factuality, linguistic correctness, conciseness, proactivity, false rejection rate (FAR) and false acceptance rate (FAR). In order to define evaluation function Q for each of them, we used specification of worse answer for every dimension. A precise description of the evaluation guidelines was given to the judge model, along with detailed specifications for each dimension, and a gold answer. We share the prompt with our source code.

3.3 Experimental protocol

In the experimental part, we create merged models by conducting two model trainings and a single merge operation (we call it a single *run*).

The pipeline of model training and merging consists of model alignment training (ORPO, DPO, KTO), win-tie rate (WTR) measurement for each model, model merging using Linear Interpolation (LERP) and Spherical Linear Interpolation (SLERP), and win-tie rate measurement of the resulting merge.

1. **We investigated linear mode connectivity** between homogeneous LERP merges (ORPO-ORPO, DPO-DPO, KTO-KTO) and heterogeneous LERP merges (ORPO-DPO, ORPO-KTO, DPO-KTO) of two distinct models with $\lambda = 0.5$ (homogenous) and $\lambda \in \{0.25, 0.5, 0.75\}$ (heterogenous). For each model combination, we created 3 merges using diverse models (trained with different shuffling of the dataset). Each source and merged model was evaluated with our protocol (Section 3.2). For every evaluation dimension, besides win-tie-rate we used custom metric to assess the performance of merged

[8] `https://huggingface.co/meta-llama/Llama-3.1-70B`

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_19

model ($m_{1,2}$) against average result of source models (m_1 and m_2):

$$WTR_{\text{mean-diff}}(m_1, m_2) = WTR(m_{1,2}) - \frac{1}{2} \cdot (WTR(m_1) + WTR(m_2))$$

2. SLERP merging led to non-functional models when applied to all parameters. **We investigated models' parameter vectors** to find possible causes and also to better understand the influence of alignment methods on groups of parameters. We calculated the L2 norms for the parameters of the aligned models and the angles Ω for heterogeneous pairs.
3. **We examined SLERP merges** that turned out to be functional after switching to LERP for RMSNorm parameter vectors. We followed the same protocol as for LERP (1.).

4 Datasets

The train dataset [9] for preference alignment consists of more than 20,000 manually annotated preference pairs, including both safety-related and neutral topics. Three distinct annotation methods were used: (1) **rating**, where each response was evaluated according to predefined metrics (informativeness, correctness, safety, fairness, conciseness, reasoning, helpfulness), (2) **ranking**, where responses were ordered according to their quality, and (3) **dialog**, where annotators took part in interactive conversations with models and selected the best responses.

The evaluation dataset contains 181 prompt-response pairs categorized as "safe" or "unsafe". Approximately half were sourced from alignment datasets (AlpacaEval [11], CREAK [14], ECQA [1], QED [10], Toxic DPO v0.2[10], Harmful Behaviors[11], Argilla[12]) and translated into Polish, covering commonsense, explanatory, and hazardous questions. The rest includes human-annotated public affairs examples and auto-generated entries. We believe that this diverse collection of examples ensures comprehensive coverage of alignment scenarios.

5 Results and Discussion

As presented in Figure 2 and Table 1 we can conclude that despite the dissimilarities of the alignment vectors' directions (measured by the angle between them, Figure 1), heterogeneous LERP merging results in LLMs that perform well and not worse than homogeneous ones. The key observation is that linear interpolation in parameter space often results in close to linear interpolation in evaluation dimensions (ones that vary the most between alignment methods: conciseness, proactivity, and factuality). This makes LERP an effective technique to obtain a model balanced between the advantages of alignment techniques. Crucially, we did not observe a significant drop in safety metrics in any of the merged models.

[9] The dataset used in this study will be publicly released in a future publication.
[10] `https://huggingface.co/datasets/unalignment/toxic-dpo-v0.2`
[11] `https://huggingface.co/datasets/mlabonne/harmful_behaviors`
[12] `https://huggingface.co/argilla`

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_19

Table 1: Results of equally-weighted homogenous and heterogenous of **LERP** and **SLERP** soups across evaluation dimensions. Values are $\mathbf{WTR}_{\mathbf{mean\text{-}diff}}$, averaged from 3 runs. LQ – linguistic correctness, Proact. – proactivity.

			Safety	Factuality	LQ	Conciseness	Proact.	FRR	FAR	Avg.
LERP	heterogenous	**ORPO-DPO**	-0.015	-0.029	**-0.013**	-0.114	-0.198	**-0.003**	-0.025	-0.057
		ORPO-KTO	-0.006	**-0.013**	-0.019	**-0.017**	**0.068**	-0.006	0.000	**0.001**
		DPO-KTO	**-0.004**	-0.029	-0.022	-0.063	-0.222	**-0.003**	**0.006**	-0.048
		Avg.	-0.008	-0.024	-0.018	-0.065	-0.117	-0.004	-0.006	
	homogenous	**ORPO-ORPO**	-0.002	**0.031**	-0.007	0.009	0.043	0.005	**0.000**	**0.011**
		DPO-DPO	**0.000**	-0.016	0.000	**0.015**	-0.080	**0.008**	-0.006	-0.011
		KTO-KTO	-0.004	-0.019	**0.002**	0.001	**-0.013**	-0.003	-0.006	-0.006
		Avg.	-0.002	-0.001	-0.002	0.008	-0.017	0.003	-0.004	
SLERP	heterogenous	**ORPO-DPO**	**0.005**	**0.017**	0.004	0.016	0.040	-0.001	**0.019**	0.014
		ORPO-KTO	-0.005	0.016	0.002	-0.031	**0.126**	**0.002**	0.003	**0.016**
		DPO-KTO	0.002	-0.025	**0.005**	**0.085**	-0.117	-0.007	0.003	-0.008
		Avg.	0.001	0.003	0.004	0.023	0.016	-0.002	0.008	
	homogenous	**ORPO-ORPO**	**0.002**	0.005	-0.009	-0.035	**0.080**	-0.005	**0.019**	**0.008**
		DPO-DPO	-0.002	0.003	-0.017	0.007	-0.093	-0.005	-0.006	-0.016
		KTO-KTO	-0.002	**0.029**	**-0.004**	**0.014**	0.018	**0.002**	-0.006	0.007
		Avg.	-0.001	0.012	-0.010	-0.004	0.002	-0.003	0.002	

Regarding analysis of models' parameters (Figures 1), we can observe slight variation between model pairs, with DPO and KTO being the most similar, and lower similarities (such as ORPO and DPO) did not make the models incompatible for merging. Also exceptionally high norm of ORPO language modeling head (LM head) parameter vector did not seem to interfere, although we consider it interesting observation that may be attributed to using prompts from the dataset as additional learning signal (which is unique for ORPO).

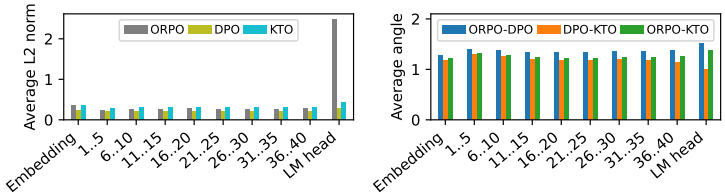

Fig. 1: Mean L2 norms (for aligned models) and angles (between models) for alignment vectors (except RMSNorm) by groups (X-axis, numbers represent transformer layers).

We observed that normalization layer parameters usually remain unchanged during alignment, except for the first transformer layer, where only ∼70% of weights match between aligned models (versus nearly 100% in other layers). The SLERP merges were functional only when RMS normalization parameters were merged with LERP, demonstrating that they may need special caution in model merging. The possible reason for this is that the outputs of normalization

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_19

layers directly affect the residual stream of models as opposed to the outputs of attention or MLP that are always followed by normalization.

We recorded a notable effect of SLERP merges on the evaluation of win-tie rates, specifically the dimensions of factuality, conciseness, and proactivity, while noting no substantial effect on safety metrics. As a result, heterogeneous ORPO-KTO but also homegeneous ORPO-ORPO merges turned out to be overall better compared to the best results of source models, with a main advantage on the proactivitv dimension.

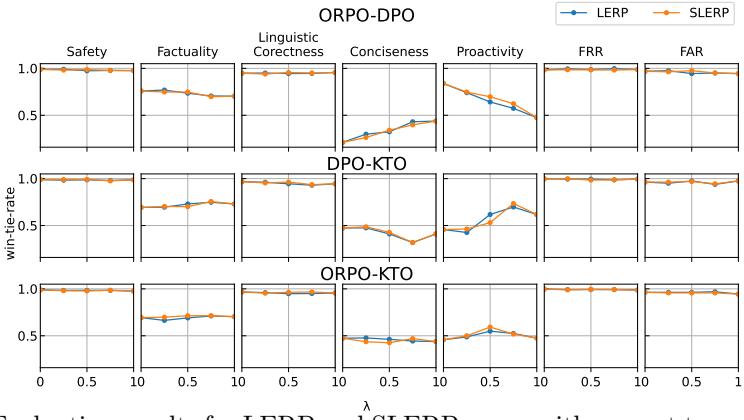

Fig. 2: Evaluation results for LERP and SLERP soups with respect to evaluation dimensions. λ controls influence from the first model, e.g. $\lambda = 0$ ORPO-DPO is a pure DPO model. Each data point is an average from 3 runs.

6 Conclusions

Our study showed that LERP and SLERP merging techniques that operate on whole model's parameter space are compatible between ORPO, DPO and KTO alignment methods utilizing various loss functions. Obtaining a balance between performance in various dimensions of large language model evaluation and models that are better on average was shown to be possible. Considering this and additional insights on alignment vectors' weights, we provided a foundation for further studies on merging aligned LLMs with more advanced techniques, focusing on dissimilarities in the alignment vectors and their connection to various dimensions of evaluation.

Acknowledgments. Financed by: (1) CLARIN ERIC (2024–2026), funded by the Polish Minister of Science (agreement no. 2024/WK/01); (2) CLARIN-PL, the European Regional Development Fund, FENG programme (FENG.02.04-IP.040004/24); (3) statutory funds of the Department of Artificial Intelligence, Wroclaw Tech; (4) the EU project "DARIAH-PL", under investment A2.4.1 of the National Recovery and Resilience Plan.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_19

References

1. Aggarwal, S., Mandowara, D., Agrawal, V., Khandelwal, D., Singla, P., Garg, D.: Explanations for CommonsenseQA: New Dataset and Models. In: ACL. "Association for Computational Linguistics" (2021)
2. Croce, F., Rebuffi, S.A., Shelhamer, E., Gowal, S.: Seasoning model soups for robustness to adversarial and natural distribution shifts (Feb 2023)
3. Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., Kiela, D.: Kto: Model alignment as prospect theoretic optimization (Feb 2024)
4. Frankle, J., Dziugaite, G.K., Roy, D.M., Carbin, M.: Linear mode connectivity and the lottery ticket hypothesis (Dec 2019)
5. Hong, J., Lee, N., Thorne, J.: Orpo: Monolithic preference optimization without reference model (Mar 2024)
6. Ilharco, G., Ribeiro, M.T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., et al.: Editing models with task arithmetic (Dec 2022)
7. Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., Wilson, A.G.: Averaging weights leads to wider optima and better generalization (Mar 2018)
8. Jang, J., Kim, S., Lin, B.Y., Wang, Y., Hessel, J., Zettlemoyer, L., et al.: Personalized soups: Personalized large language model alignment via post-hoc parameter merging (2023)
9. Jang, J., Kim, S., Ye, S., Kim, D., Logeswaran, L., Lee, M., et al.: Exploring the benefits of training expert language models over instruction tuning (Feb 2023)
10. Lamm, M., Palomaki, J., Alberti, C., Andor, D., Choi, E., Soares, L.B., et al.: Qed: A framework and dataset for explanations in question answering (2020)
11. Li, X., Zhang, T., Dubois, Y., Taori, R., Gulrajani, I., Guestrin, C., Liang, P., Hashimoto, T.B.: Alpacaeval: An automatic evaluator of instruction-following models (5 2023)
12. Mirzadeh, S.I., Farajtabar, M., Gorur, D., Pascanu, R., Ghasemzadeh, H.: Linear mode connectivity in multitask and continual learning (Oct 2020)
13. Neyshabur, B., Sedghi, H., Zhang, C.: What is being transferred in transfer learning? NeurIPS 2020 (Aug 2020)
14. Onoe, Y., Zhang, M.J., Choi, E., Durrett, G.: Creak: A dataset for commonsense reasoning over entity knowledge. OpenReview (2021)
15. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., Finn, C.: Direct preference optimization: Your language model is secretly a reward model (May 2023)
16. Rame, A., Couairon, G., Dancette, C., Gaya, J.B., Shukor, M., Soulier, L., et al.: Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In: NeurIPS (2023)
17. Ramé, A., Ahuja, K., Zhang, J., Cord, M., Bottou, L., Lopez-Paz, D.: Model ratatouille: Recycling diverse models for out-of-distribution generalization (Dec 2022)
18. Ramé, A., Ferret, J., Vieillard, N., Dadashi, R., Hussenot, L., Cedoz, P.L., et al.: Warp: On the benefits of weight averaged rewarded policies (Jun 2024)
19. Ramé, A., Kirchmeyer, M., Rahier, T., Rakotomamonjy, A., Gallinari, P., Cord, M.: Diverse weight averaging for out-of-distribution generalization (May 2022)
20. Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A.S., et al.: Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time (Mar 2022)
21. Zheng, L., Chiang, W., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) NeurIPS (2023)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_19