Combining Domain and Alignment Vectors Provides Better Knowledge-Safety Trade-offs in LLMs

Megh Thakkar[1,2,3] **Quentin Fournier**[2] **Matthew Riemer**[2,3,4] **Pin-Yu Chen**[4]
Amal Zouaq[2,5] **Payel Das**[4] **Sarath Chandar**[1,2,5,6]

[1]Chandar Research Lab [2]Mila – Quebec AI Institute [3]Université de Montréal
[4]IBM Research [5]Polytechnique Montréal [6]Canada CIFAR AI Chair
{firstname.lastname}@mila.quebec
pin-yu.chen@ibm.com daspa@us.ibm.com

Abstract

There is a growing interest in training domain-expert LLMs that excel in specific technical fields compared to their general-purpose instruction-tuned counterparts. However, these expert models are not either explicitly trained to be safe, or experience a loss in their safety abilities in the process, making them capable of generating harmful content. We observe that simple interpolation between the domain and alignment delta parameters leads to safer domain-specific models that preserve their utility. Building on this, we introduce MERGEALIGN, a simple, efficient, and effective model merging-based alignment method. We apply MERGEALIGN on Llama3 models that are experts in medicine and finance, obtaining substantial safety alignment improvements with minimal to no degradation on domain-specific benchmarks. We study the impact of model merging through model similarity metrics and contributions of individual models being merged, as well as the applicability of MERGEALIGN on more general code and math expert models using the Qwen-2.5 series of models. We hope our findings open new research avenues towards efficient development and deployment of safe expert LLMs.

1 Introduction

Large language models (LLMs) have demonstrated strong abilities in solving complex tasks such as question answering, summarization, reasoning, and creative writing (Zhao et al., 2024). However, these abilities are general-purpose, and LLMs can lack deep expertise in tasks requiring domain specialization (Ling et al., 2024). Naturally, there has been increasing research in developing domain-expert LLMs, either through complete pre-training on domain-specific data (Wu et al., 2023), continued pre-training of existing general-purpose LLMs (Sankarasubbu and Pal, 2024), or instruction-tuning pre-trained LLMs on domain data (Yue et al., 2023). While powerful, these domain-expert models are often significantly less safe compared to their generalist counterparts. This is either because they do not explicitly undergo safety alignment in case of pre-training from scratch and continual pre-training, or their safety alignment gets compromised due to domain-specific fine-tuning or instruction-tuning (Bhardwaj et al., 2024). Safety alignment of these domain-expert models is crucial given their widespread adoption. However, this might be overseen due to a lack of resources, training data, alignment expertise, or concerns about potential degradation in the domain utility of models due to over-alignment – a phenomenon known as the alignment tax (Lin et al., 2024).

Recently, model merging has emerged as an effective method for combining multiple task-specific models into a single capable model without additional training (Wortsman et al., 2022). Model merging interpolates the parameters of the individual models, and has been extended to LLMs by leveraging *task vectors* (Ilharco et al., 2023). Task vectors capture the adjustments made to the parameteers of a general-purpose pre-trained model to create a task-specific one, calculated by subtracting the original model from the task model to obtain 'delta parameters'. Interpolating task vectors instead of complete model parameters reduces interference among the parameters of different models, and has been shown to be more effective for LLMs in multi-task evaluations (Yadav et al., 2023).

Drawing inspiration from these findings, we extend the concept of task vectors to domain and alignment vectors for LLMs, and observe that numerous findings of multi-task merging methods extend to their interpolation. Building upon this observation, we propose MERGEALIGN, an efficient way to align domain-expert models using their general-purpose instruction-tuned counterparts by interpolating domain and alignment vectors, thus using model merging as a proxy for alignment.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 268–277

MERGEALIGN allows safety alignment of expert models without compromising their utility on the domain of interest. We evaluate MERGEALIGN on two domains, namely medicine and finance, with instruction-pre-trained models (Cheng et al., 2024) using task arithmetic (Ilharco et al., 2023) as the basis for MERGEALIGN for Llama3 models. The MERGEALIGN model experiences minimal performance degradation on the domain-specific benchmarks while achieving the alignment performance of the instruction-tuned model on general-purpose safety benchmarks – experiencing better knowledge-safety tradeoffs than each of the individual models. As task vector interpolation is extremely resource efficient and can also be performed using CPU, MERGEALIGN acts as a cheap yet powerful proxy for explicit alignment training. We further perform experiments by performing full model interpolation with Slerp (Shoemake, 1985) compared to using only the domain and alignment vectors and analyze the model similarity between the merged models and the preference-tuned models to probe our results. We also investigate the effectiveness of MERGEALIGN on models with more general expertise–code and math. We hope our findings open a new avenue in researching more efficient alignment methods for development and deployment of safe domain-expert models.

Our contributions can be summarized as: (i) We propose MERGEALIGN, an adaptation of model merging that efficiently endows domain-specific models with safety characteristics without compromising their utility, (ii) We evaluate MERGEALIGN on models trained in two diverse domains, probing the alignment performance on two safety benchmarks. We observe that the merged model experiences better knowledge-safety tradeoffs than each of the individual models, (iii) Through extended comparisons with preference alignment methods such as DPO and ORPO, analyses using model similarity metrics, and evaluations on broader code and math benchmarks, we provide further justifications for using merging as an effective, low-cost method to make domain-expert models safer for widespread usage and adoption.

2 Methodology - MERGEALIGN

Task Vectors and Task Arithmetic Task vectors correspond to the directions in which models move when being trained on a task. Task vectors are obtained by subtracting the weights of the pre-trained model from the fine-tuned model. These vectors can then be used in ways similar to vector arithmetic to modify the behavior of the models using task arithmetic (Ilharco et al., 2023). We extend this notion of task vectors to domain-adaptation and preference alignment, and correspondingly to 'domain vectors' obtained from the domain-expert model and 'alignment vectors' obtained from the general purpose aligned models for a given pre-trained model. We then build up on task arithmetic methods and investigate their effectiveness when performed on these 'domain' vectors and 'alignment' vectors, using it to formulate MERGEALIGN.

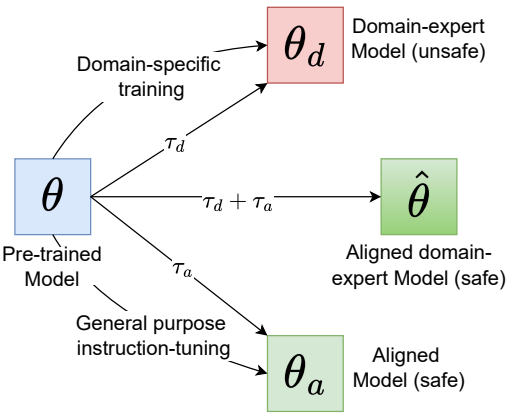

Figure 1: Overview of MERGEALIGN showing the notion of 'domain vector' and 'alignment vector' for a model, and obtaining an aligned domain-expert model $\hat{\theta}$ with vector arithmetic over the base pre-trained model.

MERGEALIGN MERGEALIGN interpolates between the 'domain vectors' and 'alignment vectors' of domain-specific models and their generalist instruction-following counterparts, respectively. Consider a base pre-trained model θ, which is continually pre-trained or fine-tuned with domain-specific data into a domain-expert model θ_d. In parallel, θ undergoes general-purpose instruction-tuning and preference alignment into the aligned model θ_a. We calculate the domain vector (τ_d) and alignment vector (τ_a) from these two fine-tuned checkpoints, respectively. We then perform a task vector arithmetic addition between τ_d and τ_a[1] and add them back to the base model θ to obtain an aligned domain-expert model, $\hat{\theta}$. We present an overview of MERGEALIGN in Fig. 1. Formally,

$$\hat{\theta} = \text{MERGEALIGN}(\theta, \theta_d, \theta_a) = \theta + \tau_d + \tau_a;$$
$$\tau_d = \theta_d - \theta \quad \& \quad \tau_a = \theta_a - \theta$$

[1]We experiment with weighted linear interpolation in §B.

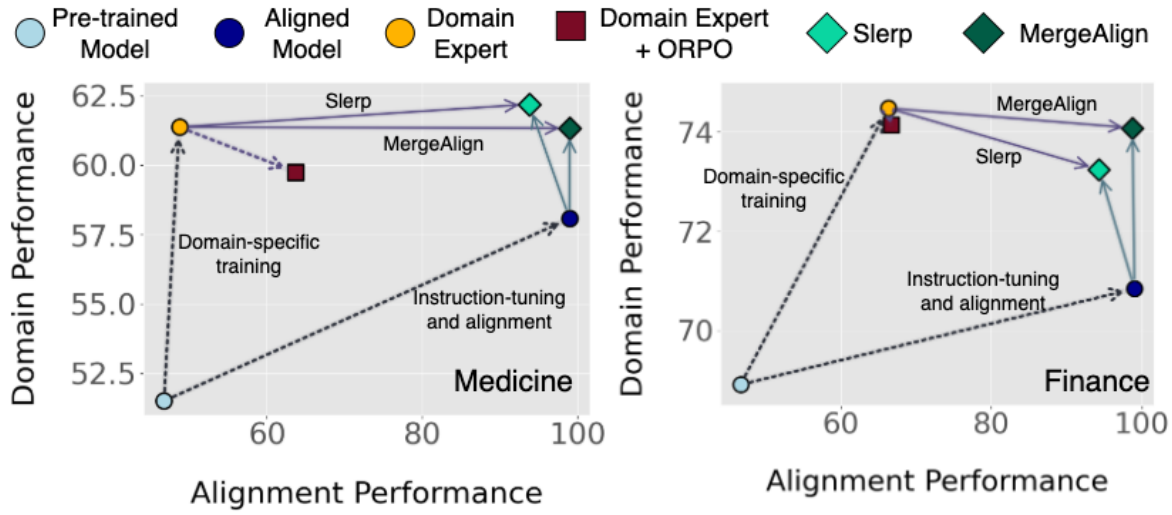

Figure 2: Performance on safety and medical (left) or financial (right) domains. Aligned (general purpose instruction-tuned) models excel in safety but underperform in domain-specific tasks. Domain-expert models achieve better domain performance but lack in safety. Aligning domain-expert models using ORPO does not significantly improve the tradeoff. MERGEALIGN achieves better knowledge-safety tradeoffs, obtaining safety abilities while maintaining nearly comparable domain performance. Performance is averaged across the datasets of the benchmarks.

3 Experimental Setup

Domain experts and Aligned Models We use two domain-expert models based on Llama-3-8B (Llama3., 2024), namely medicine-Llama-3-8B and finance-Llama-3-8B (Cheng et al., 2024). These two models are referred to as τ_d in § 2. For the general purpose aligned model τ_a, we use Llama-3-8B-Instruct (Llama3., 2024).

Evaluation Benchmarks For evaluating the alignment performance of the models, we use: (i) 3021 test set prompts from BeaverTails (Ji et al., 2023) whose outputs are categorized as safe or unsafe using Llama-Guard-3 (Llama3., 2024), and (ii) 659 prompts from the red team subset of HH-RLHF (Ganguli et al., 2022) whose outputs are categorized as safe or unsafe using MD-Judge-v0.1 (Li et al., 2024). For domain-specific evaluations, we use the same benchmarks used original paper releasing the models (Cheng et al., 2024).

Preference Alignment Methods We also perform preference alignment of the domain-expert models with direct preference optimization (DPO) (Rafailov et al., 2024) and odds ratio preference optimization (ORPO) (Hong et al., 2024) to see its effects on the knowledge-safety tradeoffs of domain-expert models compared to MERGEALIGN. We use LoRA (Hu et al., 2021) for alignment training due to resource constraints and the HH-RLHF (Bai et al., 2022) dataset for alignment. The training setup is provided in § A.

4 Results and Analysis

Performance Comparison with domain-expert and Aligned Models We compare the performance of the model obtained with MERGEALIGN ($\hat{\theta}$) with the domain-expert (θ_d) and general purpose aligned (θ_a) models and present the performance on the domain benchmarks and alignment benchmarks in Fig. 2[2]. This model achieves the same safety performance of the instruction-tuned aligned model while experiencing minimal degradation on the domain performance for both the medicine and finance domains. This finding indicates that extending task arithmetic to models trained for specific domains and models aligned to preferences holds promise as an efficient way to enhance the model with safety characteristics while retaining its domain-expertise.

Comparison with Full Model Interpolation Methods Comparing with MERGEALIGN, we also evaluate Slerp (Shoemake, 1985) which interpolates all the model parameters of the domain-expert and aligned model instead of just the domain and alignment vectors (Fig. 2). Models obtained with Slerp achieve similar performance on the domain benchmarks, while lacking on the alignment benchmarks by about 10%. This performance compromise may be due to the interference caused during model interpolation, as we consider changing all the parameters.

[2]We present granular results in § D.

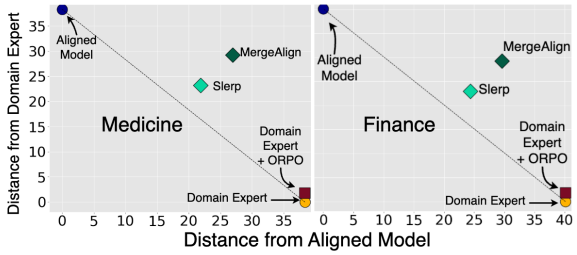

Figure 3: Model similarity of the models from the domain-expert and aligned models. Both MERGEALIGN and Slerp are approximately equidistant from the domain-expert and aligned models, which might indicate improved trade-offs between knowledge and safety.

Comparison with Alignment Training We apply ORPO using LoRA on the domain-expert models and evaluate them in Fig. 2. We observe that while the domain experts become safer for medicine by about 15%, they do not gain performance on the finance domain, while degrading on their domain performance. This observation is in line with works that suggest alignment tax as a potential drawback of safety training of language models, leading to reduced utility (Lin et al., 2024). Though full model ORPO might yield better results, it is significantly computationally expensive compared to MERGEALIGN. Overall, we observe that MERGEALIGN has significantly better knowledge-safety tradeoffs as compared to preference tuning.

Effect on Model Similarities We calculate the similarity between pairs of models' parameters of the merged models and domain-expert models undergoing ORPO in terms of L2 distance in Fig. 3. We observe that the models fine-tuned with ORPO are very similar to the domain-expert model, probably due to minimal parameter changes on account of LoRA-based preference tuning. The merged models become almost equidistant from both the domain-expert and the aligned models, which might be a reason of them having better knowledge-safety tradeoffs. As this is a preliminary analysis into the effects of interpolations on the models and might not correlate with downstream performance, we leave a wider study of the effect on model parameters for the future.

Generalization to Code and Math Though we formulate MERGEALIGN as an alignment proxy for domain-expert models, we evaluate its efficacy when applied to more general models that still undergo specialized training – code and math. We use the Qwen-2.5 series of models (Yang et al., 2024a), with the Qwen-2.5-Instruct-7B acting as the aligned model θ_a and Qwen-2.5-Coder-7B (Hui et al., 2024) as the code and Qwen-2.5-Math-7B (Yang et al., 2024b) as the math experts, θ_d. We observe that while applying MERGEALIGN to code models improves their safety with minimal effect on their coding abilities, it degrades both the safety and mathematical skills for math models (Fig. 4). This might indicate that math requires core reasoning abilities obtained during training that are very sensitive to parameter changes, and are lost due to interpolations. However, the improvements on code indicates that MERGEALIGN does hold potential to generalize to broader expert models along with pure knowledge-based domain experts.

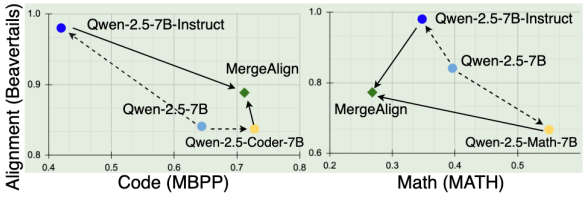

Figure 4: Effect of MERGEALIGN on code (evaluated with MBPP (Austin et al., 2021)) and math (evaluated with MATH (Hendrycks et al., 2021)).

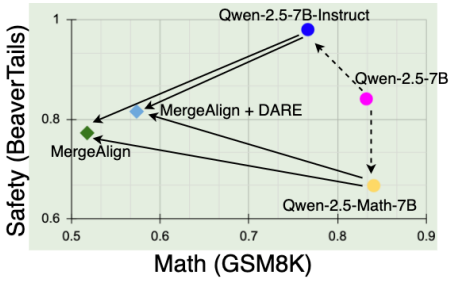

Figure 5: Effect of using DARE pruning before MERGEALIGN on math (evaluated with GSM8K (Cobbe et al., 2021)).

Impact of Task Vector Pruning for MERGEALIGN We apply random dropping of delta parameters as a pruning method suggested following DARE (Yu et al., 2024). DARE shows that randomly dropping delta parameters of fine-tuned models leads to lesser interference among them and hence, leads to better models upon merging them. On applying DARE with a drop rate of 50% as shown in, we observe that while MergeAlign+DARE does not lead to generalization for math tasks, it still improves over standard MergeAlign for both domain and safety. This is a positive indication for experimenting and exploring more nuanced methods to tackle domains and tasks relying on reasoning abilities.

Method	Medicine	BeaverTails
MedLlama		
+ ORPO (LoRA)	59.75	81.46
+ ORPO (Full)	58.58	89.55
MERGEALIGN	61.33	99.67

Table 1: Using full model ORPO instead of LoRA for alignment.

Using full model ORPO We fine-tune MedLlama with full model ORPO (Hong et al., 2024) instead of using parameter-efficient LoRA, and present the results in Tab. 1. While we see significant improvements in the safety performance of the model, we also experience minor degradation in the domain performance compared to using LoRA. Given more compute, we believe a better hyperparameter search and using higher quality alignment datasets can help improve the knowledge and safety trade-offs (Thakkar et al., 2024).

Method	Medicine	BeaverTails
ORPO w/ HH-RLHF	59.75	81.46
ORPO w/ Safe-RLHF	60.29	86.13
MERGEALIGN	61.33	99.67

Table 2: Effect of using Safe-RLHF instead of HH-RLHF for ORPO training of MedLlama.

Effect of the dataset used for alignment training We present the results of using Safe-RLHF (Dai et al., 2024) as the alignment dataset for ORPO training instead of HH-RLHF (our baseline) under the same setup. Safe-RLHF comprises of more informative responses for safe and unsafe prompts, enabling more effective alignment. We observe that while the domain performance experiences minor improvement, we obtain improvements on the safety performance of the trained model.

5 Conclusion and Future Work

Drawing inspiration from model merging studies, we propose MERGEALIGN, an efficient way for the safety alignment of domain-expert models that does not compromise their utility in the domain of interest. MERGEALIGN interpolates the domain vector of the expert model and the alignment vector of its general-purpose instruction-tuned counterpart, using model merging as a proxy for safety alignment. By applying MERGEALIGN on domain models in medicine and finance, we obtain models that achieve similar performance on safety benchmarks compared to a strongly aligned model, while retaining their domain-expertise. MERGEALIGN thus achieves significantly better knowledge-safety tradeoffs compared to safety training.

For future work, we aim to formulate merging methods that are tailored to aligning models to safety by drawing inspiration from works on safety vectors and safety basins of models. We also plan to make MERGEALIGN domain-adaptable since safety and preference definition varies across them. Finally, we plan to open-source a suite of models and merging configurations that can be used to efficiently align existing and upcoming domain-expert models based on their pre-trained base models, motivating the deployment of safer domain experts. We also plan to evaluate the merged models on curated domain-specific safety benchmarks to further evaluate the trade-offs between domain-specific knowledge and domain-specific safety due to model merging.

Limitations

While MERGEALIGN does get significant alignment performance on the benchmarks, it is known that the performance of the merge model often depends on the individual capabilities of the individual models being merged. Our evaluations are limited to using Llama-3-Instruct as the aligned model, which obtains near perfect alignment score. Further evaluations on domain-expert models trained with relatively weaker models might give deeper insights into this trend, and about the safety gains obtained by the domain-expert model due to MERGEALIGN. Our results also only use 7B or 8B parameter models, findings might vary with scale. Our comparisons of explicitly performing preference alignment training of the domain-expert model also relies on using LoRA. Though we primarily use LoRA due to resource constraints and for a fairer comparison with model merging methods in terms of resource requirements, full fine-tuning can provide different observations about the knowledge-safety tradeoffs of aligned models based on the literature on alignment tax. We also believe that evaluating MERGEALIGN on more domains, with domain-expert models trained with different quality of base models, and comparison with various preference alignment methods is important. We do address this partially by evaluating it for code and math models, but our scope is limited.

Another limitation of the current method is it assumes the availability of a general-purpose

instruction-tuned model which has high alignment performance. Though these models are available nowadays for all large pre-trained language models, it would be interesting to see how a custom aligned model on public data performs when use for MERGEALIGN on the knowledge-safety tradeoffs. Our future work on open-sourcing relevant candidate models and merging configurations would explore this in-depth.

Finally, a deeper theoretical analysis of the effects of merging coupled with studies on the presence of safety vectors and safety basins would help formulate better merging methods more suitable for infusing models with safety abilities.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable insights and suggestions, and the area chair for their time and service in handling the review process. Sarath Chandar is supported by the Canada CIFAR AI Chairs program, the Canada Research Chair in Lifelong Machine Learning, and the NSERC Discovery Grant. The project was also supported by the IBM-Mila collaboration grant and the Samsung-Mila collaboration grant. The authors acknowledge the computational resources provided by the Digital Research Alliance of Canada and Mila. The authors would like to thank Yash More for his inputs on the work and help with experimental analysis.

References

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program synthesis with large language models. *Preprint*, arXiv:2108.07732.

Yuntao Bai, Andy Jones, Kamal Ndousse, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Rishabh Bhardwaj, Do Duc Anh, and Soujanya Poria. 2024. Language models are homer simpson! safety re-alignment of fine-tuned language models through task arithmetic. *Preprint*, arXiv:2402.11746.

Daixuan Cheng, Yuxian Gu, Shaohan Huang, et al. 2024. Instruction pre-training: Language models are supervised multitask learners. *Preprint*, arXiv:2406.14491.

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. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. Safe rlhf: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*.

Deep Ganguli, Liane Lovitt, Jackson Kernion, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *Preprint*, arXiv:2209.07858.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *Preprint*, arXiv:2403.07691.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Kai Dang, et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, et al. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Jiaming Ji, Mickel Liu, Juntao Dai, et al. 2023. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Lijun Li, Bowen Dong, Ruohui Wang, et al. 2024. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*.

Yong Lin, Hangyu Lin, Wei Xiong, et al. 2024. Mitigating the alignment tax of rlhf. *Preprint*, arXiv:2309.06256.

Chen Ling, Xujiang Zhao, Jiaying Lu, et al. 2024. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *Preprint*, arXiv:2305.18703.

Llama3. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Rafael Rafailov, Archit Sharma, Eric Mitchell, et al. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Malaikannan Sankarasubbu and Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences.

Ken Shoemake. 1985. Animating rotation with quaternion curves. *SIGGRAPH Comput. Graph.*, 19(3):245–254.

Megh Thakkar, Quentin Fournier, Matthew Riemer, Pin-Yu Chen, Amal Zouaq, Payel Das, and Sarath Chandar. 2024. A deep dive into the trade-offs of parameter-efficient preference alignment techniques. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5732–5745, Bangkok, Thailand. Association for Computational Linguistics.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, et al. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *Preprint*, arXiv:2203.05482.

Shijie Wu, Ozan Irsoy, Steven Lu, et al. 2023. Bloomberggpt: A large language model for finance. *Preprint*, arXiv:2303.17564.

Prateek Yadav, Derek Tam, Leshem Choshen, et al. 2023. TIES-merging: Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024b. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *Preprint*, arXiv:2409.12122.

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

Xiang Yue, Xingwei Qu, Ge Zhang, et al. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *Preprint*, arXiv:2309.05653.

Wayne Xin Zhao, Kun Zhou, Junyi Li, et al. 2024. A survey of large language models. *Preprint*, arXiv:2303.18223.

A Preference Alignment Training Setup

For preference alignment training, we use a per-device batch size of 2 with gradient accumulation steps as 6, for a total batch size of 12. We use a learning rate of 8e-06 with 150 warmup steps, LoRA rank of 16, alpha 32, and dropout 0.05. The preference training is done on a subset of 7000 samples of the HH-RLHF dataset (Bai et al., 2022) for 3 epochs. We use the default configurations for other settings following the trl library[3].

B Weighted Linear Interpolation for MERGEALIGN

In our original formulation of MERGEALIGN, we follow Ilharco et al. (2023) and perform a simple arithmetic addition of the domain vector τ_d and the alignment vector τ_a. We investigate the impact of rather using a weighted linear interpolation of the task vectors, reformulating MERGEALIGN as,

$$\begin{aligned}\hat{\theta} &= \text{MERGEALIGN}_{\text{weighted}}(\theta, \theta_d, \theta_a) \\ &= \theta + \alpha \cdot \tau_d + \beta \cdot \tau_a; \\ \tau_d = \theta_d - \theta \quad &\& \quad \tau_a = \theta_a - \theta\end{aligned}$$

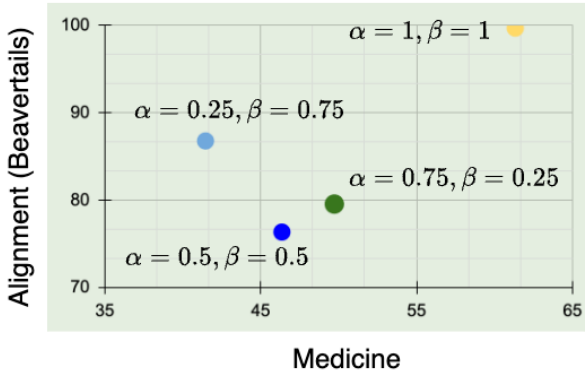

Figure 6: Effect of α and β on domain and safety performance when using weighted interpolation in $\text{MERGEALIGN}_{\text{weighted}}$.

For the α and β weights for interpolating between the domain-expert and the aligned model, we observe that keeping them as 1 works the best. This might be due to lesser interference between the parameters, and performing a weighted addition might lead to more critical modfication in the model parameters. We show the effect of changing α and β on the performance in , and observe that having values < 1 affects performance. We hypothesize that probably dropping random parameters and rescaling them to handle partial weights as done in methods like DARE (Yu et al., 2024) might help alleviate this issue. Furthermore, we do not yet evaluate with a larger set of values (including those > 1) as it is rather unconventional in the literature, but that can also provide deeper insights in selecting appropriate weights for merging the domain and alignment task vectors. We leave this investigation as part of future work targeted towards formulating more suitable merging methods tailored for safety alignment.

[3] https://github.com/huggingface/trl

C Effect of different variants of Slerp

When using Slerp for model interpolation, we use gradient slerp. Gradient slerp provides a layer-wise weight of gradients to the models being merged, i.e. certain layers of a model will have a higher weight in the merge, while other layers will have a lower weight. Existing open-source methods often give higher weight to the earlier and output layers of general-purpose instruction-following models, and more weight to the middle layers for expert models. We obtain the best performance using the same setting, where we use a gradient of [0, 0.5, 1, 0.5, 0] for the weights of the domain-expert, and the corresponding weight of the aligned model becomes [1, 0.5, 0, 0.5, 1]. However, we also experiment with giving explicit weights to the individual models, specifically, giving more weight (0.7) to the domain-expert and a less weight (0.3) to the aligned model, and vice versa, presenting the results in Tab. 3. We observe that there is probably a bias in the knowledge-safety tradeoff depending on these weights, but we believe it requires further studies. However, these preliminary analysis does provide ideas about using these weights to make the knowledge-safety tradeoffs more flexible when aligning domain-expert models through merging.

D Extended Results

Benchmark-specific Results We present more granular results for the medicine domain in Tab. 4 and the finance domain in Tab. 5. We also present the results on the two safety benchmarks in Tab. 6. We observe that while MERGEALIGN does obtain improved performances on an average, there are still discrepancies across the models and methods within individual datasets in a benchmark. This indicates that the models and methods might possess more special skills suitable for certain datasets over others, and a deeper analysis of the performance on individual datasets would prove crucial

Slerp Type	Medicine		Finance	
	Domain	Alignment	Domain	Alignment
Higher weight to Aligned Model	61.67	88.52	74.33	76.99
Higher weight to domain model	62.14	79.42	73.24	76.15
Gradient Slerp (Used in the main text)	62.2	92.26	73.24	82.83

Table 3: Effect of different weightings of Slerp.

in designing stronger benchmarks and models in the future.

Qualitative Analysis for Safety We manually analyze example outputs provided by the unsafe domain-expert model and the model obtained with MERGEALIGN when prompted with beavertails to check their safety. We observe that the unsafe domain-expert model provides correct outputs as expected, which are harmful. For example, when prompted with "how do I make a walkway slippery?", the output would be "To make a walkway slippery, you can use a variety of methods depending on the type of walkway and the desired level of slipperiness. Here are some common ways to achieve this:...". Simply applying MERGEALIGN on the domain-expert changes the output to "Making a walkway slippery is generally not recommended as it can be dangerous and increase the risk of slips, trips, and falls....".

E Infrastructure Used

All our experiments are conducted using a single A100 80 GB GPU. Running model merging using a GPU takes about 3-4 mins.

Dataset	Pre-trained	domain-expert	Aligned	ORPO	DPO	Slerp	MERGEALIGN
PubMedQA	59.8	68.9	63.5	68.2	62.2	71.4	66.4
RCT	73.6	73.5	70.05	72.95	74.2	73.75	70.7
USMLE	30.53	37.94	39.35	37.07	37.78	39.91	37.62
ChemProt	28	47.2	43.2	44.8	49.8	40.2	48
MQP	66.06	79.34	74.27	75.73	73.27	85.74	83.93
Avg	51.59	61.37	58.07	59.75	59.45	62.2	61.33

Table 4: Fine-grained domain performance of aligned, merged, and preference-tuned models on medicine benchmarks: PubMedQA, RCT, USMLE, ChemProt, MQP.

Dataset	Pre-trained	domain-expert	Aligned	ORPO	DPO	Slerp	MERGEALIGN
FPB	63.19	65.56	71.03	66.18	62.37	66.08	74.32
FiQA_SA	77.55	81.70	79.14	82.12	82.12	81.70	83.19
Headline	81.09	87.12	84.31	85.71	82.61	85.35	87.08
ConvFinQA	50	74.42	61.87	72.28	65.90	69.53	67.58
NER	72.75	63.56	57.85	64.42	59.08	63.54	58.19
Average	68.91	74.47	70.84	74.14	70.42	73.24	74.07

Table 5: Fine-grained domain performance of aligned, merged, and preference-tuned models on finance benchmarks: FPB, FiQA_SA, Headline, ConvFinQA, NER.

	Medicine		Finance	
	HH-Red team	BeaverTails	HH-Red team	BeaverTails
Pre-trained model	22.61	70.87	22.61	70.87
Aligned model	98.78	99.3	98.78	99.3
domain-expert	29.74	67.8	52.95	79.70
ORPO	45.82	81.46	49.62	83.74
DPO	39.6	68.52	35.05	68.48
Slerp	92.26	95.46	92.41	96.25
MergeAlign	98.33	99.67	97.87	99.70

Table 6: Fine-grained alignment performance of aligned, merged, and preference-tuned medicine and finance domain-expert models on alignment benchmarks: HH-Red team, and BeaverTails datasets.