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

Megh Thakkar[1,2] **Yash More**[2] **Quentin Fournier**[1,2] **Matthew Riemer**[2,3] **Pin-Yu Chen**[3]
Amal Zouaq[2,4] **Payel Das**[3] **Sarath Chandar**[1,2,4,5]
[1]Chandar Research Lab [2]Mila – Quebec AI Institute [3]IBM Research
[4]Polytechnique Montréal [5]Canada CIFAR AI Chair
{megh.thakkar, sarath.chandar}@mila.quebec

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 often experience a loss in their safety abilities in the process, making them capable of generating harmful content. As a solution, we introduce an efficient and effective merging-based alignment method called MERGEALIGN that interpolates the domain and alignment vectors, creating safer domain-specific models while preserving their utility. We apply MERGEALIGN on Llama3 variants that are experts in medicine and finance, obtaining substantial 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. We hope our findings open new research avenues and inspire more efficient development 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 as they either 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 task-specific models into a single model without additional training [Ilharco et al., 2023]. Model merging interpolates the parameters of multiple models, and has been extended to LLMs by leveraging *task vectors*. Task vectors capture the adjustments made to the weights of a general-purpose pre-trained model to create a task-specific one, calculated by subtracting the original model from the task model. Combining them has been shown to yield minimal performance degradation in multi-task evaluations [Yadav et al., 2023].

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Drawing inspiration from these findings, we expand the concept of task vectors to domain and alignment vectors for LLMs, computed by leveraging the corresponding domain adapted and aligned models of a pre-trained model, respectively. Leveraging state-of-the-art model merging methods, we propose MERGEALIGN, an efficient way to align domain expert models using their general-purpose instruction-tuned counterparts by interpolating domain and alignment vectors. 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. We observe that the MERGEALIGN model experiences minimal performance degradation on the domain-specific benchmarks. However, it is able to achieve the alignment performance of the instruction-tuned general purpose model as evident from the evaluations using two safety benchmarks, achieved at very low cost. We further compare the merged model with the preference-tuned model, where the domain expert models undergo preference alignment training. Our evaluations show reduced knowledge-safety trade-offs in MERGEALIGN compared to preference tuning, while the former is also more cost-efficient. We further provide insights into the model similarity between the merged models and the preference-tuned models, and ablate on performing full model interpolation with Slerp [Shoemake, 1985] compared to using only the domain and alignment vectors. We hope our findings open a new avenue in researching more efficient alignment methods when dealing with domain-specific scenarios and promote development of safer models that are widely adopted for everyday uses. Our contributions can be summarized as,

- We propose MERGEALIGN, an adaptation of model merging that efficiently endows domain-specific models with safety characteristics without compromising their utility.
- We evaluate MERGEALIGN on models trained in two diverse domains, probing the alignment performance on two safety benchmarks. We observe that the merged model achieves strong alignment performance while retaining domain expertise.
- Through extended comparisons with preference alignment methods such as DPO and ORPO and analyses using model similarity metrics, we provide justification for using merging as an effective, low-cost method to make domain expert models safer for widespread 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]. Similar observations has also been made with safety vectors [Bhardwaj et al., 2024], obtained by using safety-aligned models and their unaligned counterparts. 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.

MERGEALIGN MERGEALIGN uses the corresponding '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, leading to the domain expert model θ_d. In parallel, the pre-trained model θ undergoes general-purpose instruction-tuning and preference alignment, leading to 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 linear interpolation between τ_d and τ_a with weights α and β 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 and formalize it as,

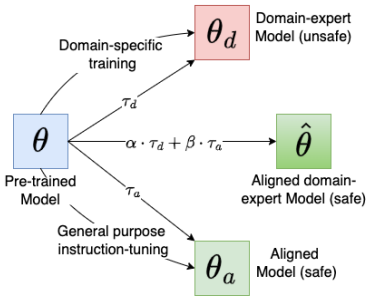

Figure 1: Overview of MERGEALIGN showing the notion of 'domain vector' and 'alignment vector' for a model, and obtaining an aligned domain expert model $\hat{\theta}$ with vector arithmetic over the base pre-trained model.

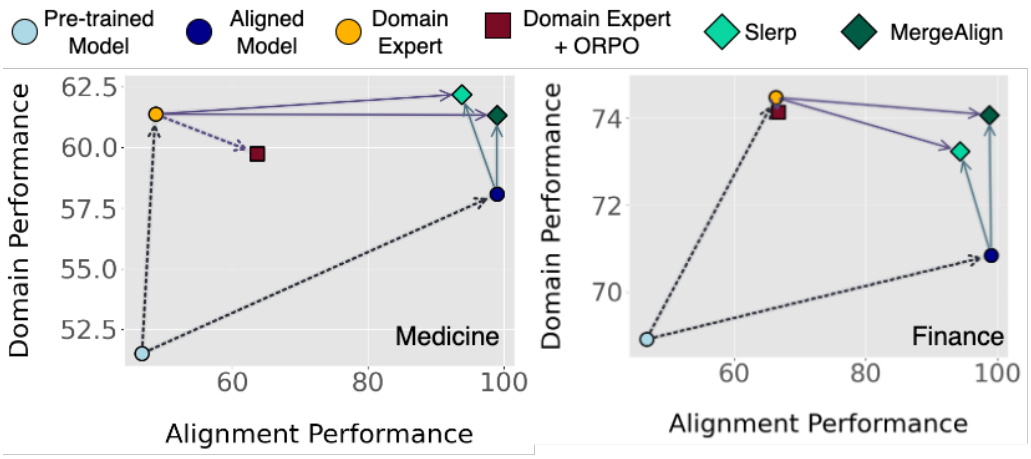

Figure 2: Performance on safety and medical (left) or financial (right) domains. Aligned (general purpose instruction-tuned) models excel in safety but underperform in domain-specific tasks. Domain expert models achieve the highest domain performance but perform the worst in safety. Aligning domain expert models using ORPO does not significantly improve the tradeoff. In contrast, MERGEALIGN with the aligned and domain expert model delivers safety levels close to instruction-tuned models while maintaining nearly comparable domain performance. Performance is averaged across the datasets of the benchmarks.

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

3 Experimental Setup

Domain expert and Aligned Models We apply MERGEALIGN on models of Llama-3-8B [et al., 2024] in two speciality domains, 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 alignment model τ_a, we use Llama-3-8B-Instruct, as it has been instruction-tuned from the same base checkpoint and carries the sufficient linear mode connectivity conditions suitable for model merging [Wortsman et al., 2022].

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 [et al., 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 by the original domain expert models.

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] using the HH-RLHF [Bai et al., 2022] dataset to see its effects on the knowledge-safety tradeoffs when trained explicitly on human preferences. We use LoRA [Hu et al., 2021] for alignment training due to resource constraints. The training setup is provided in § 6.1.

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[1]. 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.

[1]We present more granular benchmark specific results in § 6.4.

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 in 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 can may be due to the interference caused during model interpolation, as we consider changing all the parameters. These results indicate that merging the domain and alignment vectors is more beneficial, but model interpolation can also help. We plan to further study the effects of various merging methods as future work.

Comparison with Preference Alignment Methods We perform preference alignment of the domain expert models and evaluate them in Fig. 2. We observe that while the domain expert models gain better safety performance due to preference alignment for medicine by about 15%, they do not gain performance on the finance domain, and suffer on the domain performance for both domains. This might be either a characteristic preference alignment, in line with works that suggest alignment tax as a potential drawback of safety training of language models leading to decreased utility [Lin et al., 2024] or due to using parameter-efficient training, which still requires more compute than model merging. Overall, we observe that MERGEALIGN has significantly better knowledge-safety tradeoffs as compared to preference tuning of domain expert models.

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 aligned with preference data from the domain expert models still are very similar to the domain-expert model. This might be due to minimal parameters being changed due to LoRA-based preference tuning. However, the merged model become almost equidistant from both the domain expert and the general purpose models, which might be a reason of them being safe while retaining domain expertise. We leave deeper analysis of this as a potential wider study of model merging in the future. We believe this analysis is preliminary to get more insights into the effects of interpolations on the modeel parameters, and might not directly correlate with downstream performance. However, we believe that analyzing the parameter space of these variants can lead to interesting insights.

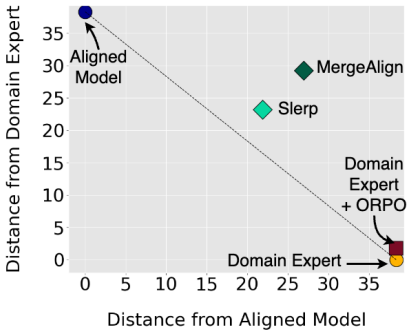

Figure 3: Model similarity of the merged and preference tuned models using L2 distance from the domain expert and aligned models (for medicine domain). Model aligned via ORPO behaves similarly to the domain model and is close to it. Both MERGEALIGN and Slerp are approximately equidistant from the domain and instruct models.

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. 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. Further analyses on performance and model similarity comparison of MERGEALIGN with preference alignment methods like DPO and ORPO justifies and validates its benefits. We hope to incentivize research into efficiently and effectively aligning domain expert models given their widespread usage in the real world.

For future work, we aim to formulate merging methods that are particularly tailored in aligning models and making them safer, for which we plan to draw inspiration from works on safety vectors and safety basins of models. We also plan to use MERGEALIGN to make alignment flexible to domains, since the definition of safety and preferences often varies across domains. The findings of model merging works on more than 2 tasks can be leveraged to perform multi-preference domain-specific alignment as well

effectively. Finally, we plan to open-source a suite of aligned models and merging configurations that can be used to efficiently align majority of existing and upcoming domain expert models based on their pre-trained base models, motivating the development of safer widely-used language models.

6 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 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 which can further provide motivations for coming up with methods that don't degrade the domain knowledge of the model. Overall, we 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. Another limitation of the current method is it assumes the availability of a general-purpose instruction-tuned model which has high alignment performance. Though this is 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.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Rishabh Bhardwaj, Do Duc Anh, and Soujanya Poria. Language models are homer simpson! safety re-alignment of fine-tuned language models through task arithmetic, 2024. URL https://arxiv.org/abs/2402.11746.

Daixuan Cheng, Yuxian Gu, Shaohan Huang, et al. Instruction pre-training: Language models are supervised multitask learners, 2024. URL https://arxiv.org/abs/2406.14491.

Abhimanyu Dubey et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Deep Ganguli, Liane Lovitt, Jackson Kernion, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, 2022. URL https://arxiv.org/abs/2209.07858.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model, 2024. URL https://arxiv.org/abs/2403.07691.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, et al. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Jiaming Ji, Mickel Liu, Juntao Dai, et al. 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*, 2023. URL https://openreview.net/forum?id=g0QovXbFw3.

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

Yong Lin, Hangyu Lin, Wei Xiong, et al. Mitigating the alignment tax of rlhf, 2024. URL https://arxiv.org/abs/2309.06256.

Chen Ling, Xujiang Zhao, Jiaying Lu, et al. Domain specialization as the key to make large language models disruptive: A comprehensive survey, 2024. URL https://arxiv.org/abs/2305.18703.

Rafael Rafailov, Archit Sharma, Eric Mitchell, et al. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Malaikannan Sankarasubbu and Ankit Pal. Openbiollms: Advancing open-source large language models for healthcare and life sciences, 2024. URL https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

Ken Shoemake. Animating rotation with quaternion curves. *SIGGRAPH Comput. Graph.*, 19(3): 245–254, July 1985. ISSN 0097-8930. doi: 10.1145/325165.325242. URL https://doi.org/10.1145/325165.325242.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, 2022. URL https://arxiv.org/abs/2203.05482.

Shijie Wu, Ozan Irsoy, Steven Lu, et al. Bloomberggpt: A large language model for finance, 2023. URL https://arxiv.org/abs/2303.17564.

Prateek Yadav, Derek Tam, Leshem Choshen, et al. TIES-merging: Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=xtaX3WyCj1.

Xiang Yue, Xingwei Qu, Ge Zhang, et al. Mammoth: Building math generalist models through hybrid instruction tuning, 2023. URL https://arxiv.org/abs/2309.05653.

Wayne Xin Zhao, Kun Zhou, Junyi Li, et al. A survey of large language models, 2024. URL https://arxiv.org/abs/2303.18223.

Supplementary Material

6.1 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[2].

6.2 Hyperparameters for MergeAlign

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.

6.3 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. 1. 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.

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 1: Effect of different weightings of Slerp.

6.4 Extended Results

We present the dataset-wise results for the medicine domain (Tab. 2) and the finance domain (Tab. 3) for the different settings: pre-trained, domain expert, aligned, preference-tuned, and merged models. We also present the results on the two safety benchmarks in Tab. 4.

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

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	74.268	75.73	73.27	85.74	83.93
Avg	51.59	61.37	58.07	59.75	59.45	62.2	61.33

Table 2: 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 3: 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 4: 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.