MEAV: Model Editing with Alignment Vectors for inference time LLM alignment in single and multidomain preference spectrum

Anonymous ACL submission

Abstract

Aligning Large Language Models (LLM) to address subjectivity and nuanced preference levels requires adequate flexibility and control, which can be a resource-intensive and time-consuming procedure. Existing training-time alignment methods require full re-training when a change is needed and inference-time ones typically require access to the reward model at each inference step. We introduce **MEAV**, an inference-time model-editing-based LLM alignment method that learns encoded representations of preference dimensions, called *Alignment Vectors* (AV). These representations enable dynamic adjusting of the model behavior during inference through simple linear operations. Here, we focus on three gradual response levels across three specialized domains: medical, legal, and financial, exemplifying its practical potential. This new alignment paradigm introduces adjustable preference knobs during inference, allowing users to tailor their LLM outputs while reducing the inference cost by half compared to the prompt engineering approach. Additionally, we find that AVs are transferable across different fine-tuning stages of the same model, demonstrating their flexibility. AVs also facilitate multidomain, diverse preference alignment, making the process 12x faster than the retraining approach.

1 Introduction

Aligning LLMs is crucial for adapting them to meet human preferences. Standard training-time alignment methods, such as RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2024), are conducted during model training. However, making nuanced preference adjustments during inference with these approaches necessitates retraining, which requires substantial amounts of time, preference data and computational resources. Inference-time LLM alignment, by contrast, delays the alignment process until inference (Wang et al., 2024). While preference alignment can be achieved through training-time methods or targeted prompting, fine-grained control over preferences at inference remains largely unexplored in current State-of-the-Art (SOTA) works (Sahoo et al., 2024; Guo et al., 2024). This research introduces an inference-time model editing technique via *Alignment Vectors* (AV), offering users dynamic preference adjustments without additional computational overhead.

Due to their extensive capabilities, LLMs are now employed in different fields, but the diverse needs of a broad customer base require that LLM outputs be carefully refined. For instance, while a healthcare provider might need detailed medical responses for professional use, a public health forum may prefer more generalized information to avoid misinterpretation. Although prompt engineering can temporarily address these needs, it becomes costly when scaled (Li et al., 2023).

Furthermore, managing multiple alignment objectives can be complex. Consider an insurance company that needs expert legal responses, generic financial answers, and to avoid medical responses; balancing these demands poses a significant challenge. A joint training with targeted preference levels can resolve the problem, however, it lacks flexibility, and training can be resource intensive. Hence, at present, there is no work that addresses such preference flexibility in the inference time. Thus, developing flexible, inference-time adjustable model alignment to manage costs and maintain efficiency in the long term remains a major research gap.

Why not the conventional training-time approach? In contrast to conventional approaches, inference time alignment provides flexibility and adaptability by enabling dynamic adjustments to model behavior based on task or user needs without retraining.

Preference dimensions like helpfulness, harmlessness, and honesty are well-studied, with some work exploring their controllability via numerical levels (Bai et al., 2022; Ji et al., 2024; Guo et al., 2024). However, specialized dimensions offer finer granularity, enabling better control during inference. To enhance preference tunability, we focus on proficiency levels in specialized domains while also demonstrating tunability in a general domain, such as safety. Since existing literature lacks domain-specific preference alignment datasets, we generate synthetic Query-Response pairs by deriving queries from the PersonaHub dataset (Chan et al., 2024) and augmenting them with novel personas created via LLM-generated prompts.

In addition, to achieve inference time preference tunability, we propose a simple technique called **M**odel **E**diting via **A**lignment **V**ector (MEAV), which is based on the concept of *Task Arithmetic* (Ilharco et al., 2023). AVs can be obtained by directly subtracting the base model parameters from the aligned model, and can be added in the inference time. Hence, our first research question (**RQ1**) Are alignment vectors valid representation of the preference dimensions? To address this question, we systematically integrate the alignment vector into the base model with varying weights, both positive and negative, and analyze the resulting changes in model behavior. Our second research question is posed as (**RQ2**) Can we calibrate different alignment vectors to achieve diverse multi-domain preference? We address RQ2 through different domain-specific AV-integration strategy.

The key contribution of this work are:

- We frame LLM alignment in single and multiple domains as a model editing problem and introduce an inference-time tunable mechanism, which allows flexible adjustment of generation output along the preference dimension.
- We generate a synthetic dataset with a total of 38k queries, each paired with responses categorized into three levels of specialized subject matter proficiency across three specialized domains: Medical, Financial, and Legal. The dataset will be available through this link.
- By adjusting the merging coefficients, we achieve diverse, multidomain behaviors efficiently, saving time and resources. Unlike joint training, which requires p^D adjustments for D domains and p preference levels, our method only requires D training runs, reducing resource usage by a factor of p^D/D.

2 Related Works

Prompt engineering echniques, such as zero-shot, few-shot, and Chain-of-Thought (COT) prompting have proven effective in aligning language model responses to user queries during inference time (Radford et al., 2019; Sahoo et al., 2024; Wei et al., 2022). However, it comes with expensive inference time and cost when scaled. Additionally, effective prompt engineering assumes that the user is skilled at interacting with LLMs (Meskó, 2023; Oppenlaender et al., 2023).

Li et al. introduced Inference-Time Intervention (ITI), which identifies a sparse set of attention heads with high linear probing accuracy for a target task and shifts their activation along task-correlated directions during inference time (Li et al., 2024). A similar approach was eplored to learning Safety Related Vectors (SRV), to steer harmful model outputs towards safer alternatives (Wang et al., 2024). However, these methods were target domain-specific and not controllable. Huang et al. introduced DeAl, an alignment method that treats alignment as a heuristic-guided search process (Huang et al., 2024). Liu et al. studied regularization strength between aligned and unaligned models to have control over generation (Liu et al., 2024). Although closely related to our work, their method lacks clarity on whether fine-grained preference levels can be achieved. Researchers controlled attributes of generated contents by adding control token in the prompt (Guo et al., 2024; Dong et al., 2023). Despite its effectiveness, this method requires training LLMs with a particular data format, which restricts the flexibility of control during inference.

Rame et al.'s work is closely related to our multi-domain preference alignment (Rame et al., 2024). However, their approach focuses on training-time alignment by interpolating weights from models fine-tuned on diverse rewards to achieve Pareto-optimality. In contrast, our work introduces a preference adjustment strategy that operates at inference time, in addition to achieving multi-dimensional alignment. Similarly, while Jang et al. address personalized preference alignment and post-hoc merging, our approach provides a unique capability: preference level adjustment (Jang et al., 2023).

3 Methodology

MEAV starts with deriving the AVs, followed by the dynamic weighted integration of these AVs with the unaligned model.

3.1 Obtaining Alignment Vector

To obtain the AVs, we first perform alignment through the DPO algorithm, using an 'ipo' loss function to create a domain-specific aligned model (Rafailov et al., 2024; Azar et al., 2024). We get AVs by subtracting the weights of an unaligned model from the weights of the same model after alignment on a task. If $\theta_{aligned}$ denotes the model parameter after aligning on a preference dimension, then the AV can be obtained by the following:

$$\theta_{AV} = \theta_{aligned} - \theta_{unaligned} \quad (1)$$

3.2 Single Domain Alignment

To enable preference tunability across different domains, we perform a weighted integration of the AVs into the base (or unaligned) model, where the weights can be both positive and negative. We hypothesize that this gradual integration will result in a corresponding gradual increase or decrease in the model's proficiency. This process is governed by the following equation.

$$\theta_{aligned} = \theta_{unaligned} + \lambda * \theta_{AV} \quad (2)$$

By adjusting the value of λ, we aim to control the proficiency of the model's generated responses. Assuming when $\lambda = 0$, the model remains unaltered and functions as the base, unaligned model. If the θ_{AV} encodes the expert behavior in a certain domain, as λ increases towards 1, the model becomes increasingly aligned, achieving full proficiency at $\lambda = 1$.

We further hypothesize that when λ takes on negative values, the model's behavior tends to reverse the preference ranking. For instance, if the base model typically generates generic responses and the aligned model is designed for expert-level responses, moving λ in the negative direction will shift the model towards avoidance behavior. Therefore, to control the proficiency of the responses, adjusting λ is sufficient, eliminating the need to train the model with a new preference configuration.

3.3 Multidomain Alignment

When dealing with multiple domains simultaneously, the interaction between these domains can present a significant challenge. While individual preference vector encodes domain-specific attributes, they also embed proficiency levels which can easily generalize and negatively affect multidomain diverse behavior. This complexity can make it difficult to integrate multiple domains effectively.

Our goal is to achieve a diverse multidomain preference, which we approach by using the following equation:

$$\theta_{multidom\_aligned} = \alpha\theta_{AV\_dom1} + \beta\theta_{AV\_dom2} + \gamma\theta_{AV\_dom3} \quad (3)$$

In this equation, α, β and γ represent the integration coefficients for the domains in question, respectively. By identifying different sets of these coefficients, we aim to achieve varying levels of preference across the three domains.

4 Synthesizing Specialized Preference Data

To gather data for preference tuning on response proficiency levels, we employ two methods to collect queries: "PersonaHub" (Chan et al., 2024) and "CreatePersona." Figure 1 provides a detailed overview of the process. Notably, all generated persona, queries, responses, and the prompts used are in English.

4.1 Query Generation

We initiate the generation with a hierarchical process called "CreatePersona." We begin by randomly generating a few persona-query pairs by prompting Claude-3-Sonnet (Anthropic, 2024). To preserve diversity, we limit the initial set to five pairs, as we found generating too many at the outset reduces variation. From each initial persona, we recursively generate additional persona-query pairs that are relevant to the root persona. We randomize this process three times.

To further diversify the dataset, we supplement our generated personas by randomly sampling an equal number from the PersonaHub dataset (Chan et al., 2024), licensed as cc-by-nc-sa-4.0. Using these selected personas, we prompt Claude-3-Sonnet (Anthropic, 2024) to generate specialized domain queries.

We chose Claude-3-Sonnet over GPT-4 for two main reasons: First, Claude-3-Sonnet has consistently demonstrated performance on par with GPT-4, often ranking among the best foundational mod-

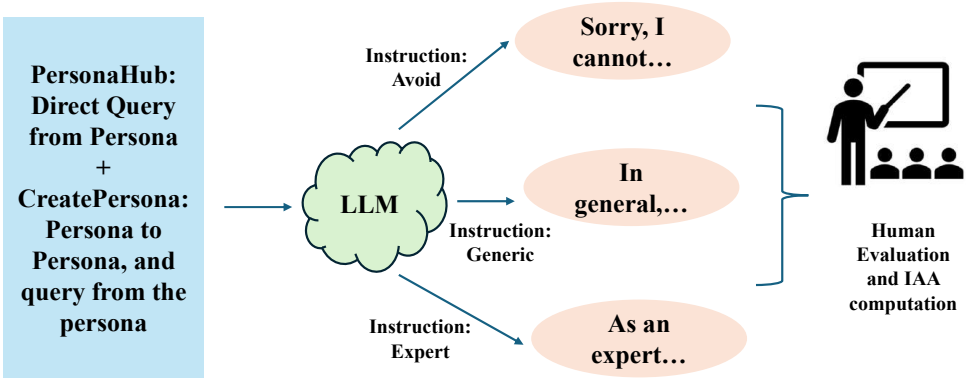


Figure 1: The process of data collection. Personas are sourced from both the PersonaHub dataset and the CreatePersona method. These personas are then fed to an LLM to generate queries. The LLM is prompted with specific instructions to produce responses across three proficiency levels. Following this, human evaluation is conducted to ensure the accuracy and quality of the generated response levels.

els. Second, we opted to use GPT-4 as an independent evaluator and sought to mitigate the known bias where evaluators tend to favor their own outputs over those generated by other models (Zheng et al., 2024; Anthropic, 2024).

After a thorough clean-up, involving truncation, and reformatting, we obtained 13,000 personas for the medical domain, 12,374 personas for the financial domain, and 12,867 personas for the legal domain. Each persona is accompanied by queries pertinent to their respective specialized domains.

4.2 Response Generation

We generate the response from the queries into three distinct levels: avoidance of response (Avd), generic response (Gen), and expert response (Exp). Detailed instructions are provided to the LLM to facilitate the generation of these responses (see Appendix C). Furthermore, we observe a progressive increase in response length from the avoidance level to the expert level. To mitigate potential bias associated with response length, we instructed the LLM to produce responses of random lengths.

4.3 Human Evaluation of multi-level response generation

To evaluate the quality of the generated responses, we conduct a small experiment involving three annotators, and compute the Inter-Annotator Agreement (IAA). Each annotator is asked to categorize a set of LLM-generated responses into one of three categories: Avd, Gen, and Exp. We provide the annotators with clear definitions of these categories. Each annotator reviews 30 queries along with their three-level responses, with at least 15 examples shared between every pair of annotators. This allows us to compute the average Cohen's kappa score, which is found to be 0.84 (Cohen, 1960), indicating substantial agreement among the annotators.

We also calculate the average annotation agreement for each annotator with the LLM generation. Responses generated with the Avoidance instruction have the fewest disagreements or misclassifications. However, some Gen and Exp responses are occasionally misclassified from one another. We observe that certain responses, although aligned with the expert spectrum, are misidentified as generic due to their tone, and vice versa. Additionally, a few avoidance responses provide basic information, leading to their misclassification as Gen responses. These findings suggest that the levels may represent a continuous spectrum rather than distinct categories, highlighting the need for further research to more precisely define these proficiency levels.

5 Experiments

5.1 Evaluation Metric

To assess the performance after alignment, we use a metric called *preference accuracy* (pref. acc). This metric reports the accuracy at each alignment level. To calculate it, we first compute the token-level mean log-probability (MLP) for each of the three response levels across all queries for the aligned model. Then, for each sample in the validation set, we determine which alignment level has the highest log-probability. For example, in proficiency level alignment, it can be among Exp, Gen, and Avd. Finally, we report the percentage of sam-

ples where each alignment level had the highest log-probability in the validation set. A higher preference accuracy in an alignment spectrum indicate the dominant behavior of that level.

To illustrate, for a query $q \in Q$, the mean log-probability for response $r \in R$, where R can be different alignment levels, is computed for model M_λ as:

$$\textit{MLP}(r, q, M_\lambda) = \frac{1}{T_r(q)} \sum_{i=1}^{T_r(q)} \log P(t_i \mid \text{ctx}, M_\lambda) \tag{4}$$

where $T_r(q)$ is the response length, t_i is the i^{th} token and ctx is the previously processed context. The preferred alignment level is:

$$r^*(q) = \arg\max_{r \in R} \textit{MLP}(r, q, M_\lambda).$$

The preference accuracy for level r is:

$$\textit{Pref. Acc}(r) = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}[r^*(q) = r],$$

where $\mathbf{1}[r^*(q) = r]$ is the indicator function. Higher $\textit{Pref. Acc}(r)$ indicates the dominant behavior of the preference alignment level r. A similar approach was also used in pairwise preference accuracy computation in (Stiennon et al., 2020).

Additionally, we use an auxiliary metric as "GPT-4 judged generation accuracy", where we generate the responses from queries in a sample, and ask GPT-4 to annotate it as one of the three levels (Zheng et al., 2024). After that, we simply report the percentage of each annotated alignment level.

5.2 Baseline Approaches

Since existing model-editing methods lack inference-time controlled alignment, we use 'prompting' as a baseline, instructing the LLM to generate responses at predefined proficiency levels. Unlike model editing, this enables discrete levels rather than a spectrum. Our second baseline, 'Joint Training,' combines multidomain data to align responses across proficiency levels, offering insights despite being a training-time method. We also report the model's 'default' performance, where queries are prompted without additional instructions or edits.

5.3 Model and Training Configuration

We define three main preference levels—"expert," "generic," and "avoidance"—for specialized domain proficiency and use DPO training with a fixed beta of 0.1, where "expert" is preferred over "generic," and "generic" over "avoidance." To demonstrate preference tunability, we vary λ in increments of 0.1, capturing significant behavioral shifts. As a base model, we use *Mistral-7B-Instruct-v0.3* (Jiang et al., 2023) (licensed apache-2.0), training on NVIDIA A100 GPUs with an 80/20 train/test split, and 3% for validation. We run one epoch at a batch size of 4 and stop training when validation loss converges.

Apart from the special domain dataset, we also use the PKU-SafeRLHF dataset (licensed cc-by-nc-4.0) for safety and helpfulness alignment experiments (Ji et al., 2024).

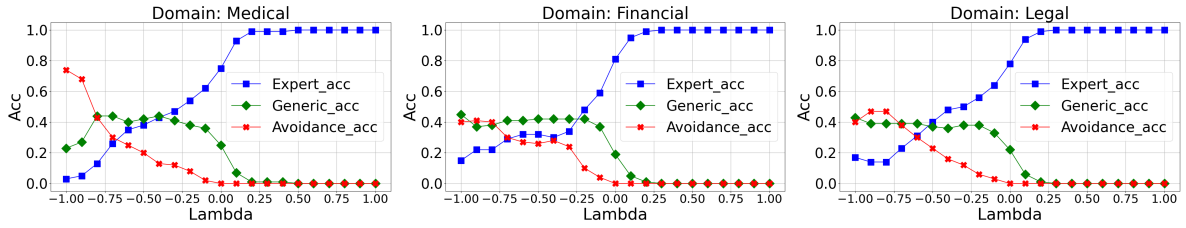
6 Results and Discussion

6.1 Single Domain Preference Tuning

We use the AV derived by aligning the model to generate responses at an expert-level within a given domain. It facilitates model editing which introduces a tunable parameter, allowing the user to control the proficiency level of the generated responses in a continuum. Consequently, one alignment vector is established for each domain, enabling the model to navigate and produce output across varying spectra of proficiency. This, in turn, also addresses **RQ1**.

Table 1 shows that simply adding instructions for specific expertise (i.e., prompting) does not significantly improve preference accuracy, while nearly doubles inference cost. Notably, the base model achieves high expert-level accuracy even with prompts from a different LLM (Claude-3-Sonnet), though it performs poorly in generic (0.31) and avoidance (0.15) categories. For MEAV, adding the AV at different λ values shifts the model's likelihood of generating expert responses: negative λ reduces expertise (with avoidance at $\lambda = -1.2$), while in the medical domain, $\lambda = -0.7$ yields generic behavior and $\lambda = 0.5$ produces full expertise.

Figure 2 illustrates the tunable nature of the preference expertise spectrum across all three domains. Notably, at $\lambda = 0$, the model predominantly generates expert responses in all domains. In the medical domain, the model reaches the higher end of the expertise spectrum when λ exceeds 0.3. Between $\lambda = -0.4$ and $\lambda = -0.8$, the model exhibits vary-



(a) (b) (c)

Figure 2: By changing the λ parameter in the MEAV process, we achieve different alignment objectives. In (a), when $\lambda > 0.3$, we find the model aligning with expert answers to medical queries by prefering expert responses over the others. However, when $\lambda < -0.8$, we see the model prefers avoidance of responses. In between these points, we observe the model answering generically to medical queries. (b) and (c) demonstrates this behavior for financial and legal domains respectively. Here λ acts as a "tunable knob", through which users can adjust the behavior of the model, and have the expertise level at any spectrum they want

Domain	**Technique**	**Target behavior**	**Pref. Acc.**			**GPT-4 judged gen. acc**		
			Exp	**Gen**	**Avd**	**Exp**	**Gen**	**Avd**
Medical	Default		.75	.25	0	.90	.05	.05
	Prompting	Exp	.78	.22	0	.90	.05	.05
		Gen	.69	.31	0	.50	.50	0
		Avd	.60	.25	.15	.15	.55	.30
	Ours: MEAV	Exp (.5)	**.95**	0	.05	**1.0**	0	0
		Gen (-.7)	.26	**.44**	.30	0	**.60**	.40
		Avd (-1.2)	.03	.13	**.84**	.05	.20	**.75**
Financial	Default		.81	.19	0	.85	.15	0
	Prompting	Exp	.84	.16	0	.95	.05	0
		Gen	.57	.43	0	.75	.25	0
		Avd	.35	.49	.16	.20	.60	.20
	Ours: MEAV	Exp (.3)	**.85**	.15	0	**1.0**	0	0
		Gen (-.4)	.30	**.42**	.28	.35	**.50**	.15
		Avd (-1.4)	.07	.20	**.73**	0	.15	**.85**
Financial	Default		.78	.22	0	.85	.15	0
	Prompting	Exp	.79	.21	0	1.0	0	0
		Gen	.59	.41	0	.65	.35	0
		Avd	.41	.30	.29	.15	.40	.45
	Ours: MEAV	Exp (.3)	**1.0**	0	0	**1.0**	0	0
		Gen (-.7)	.23	**.39**	.38	o	**.65**	.35
		Avd (-1.4)	0	.20	**.80**	0	.05	**.95**

Table 1: How MEAV performs to steer different domain expertise response level. The *Default* behavior indicates $\lambda = 0$, i.e., the model with no alignment. Tuning Lambda to different values with our MEAV approach leads to varying levels of proficiency responses. As such, we observe Exp, Gen, and Avd behavior just by aligning one model.

ing degrees of generic behavior and beyond that, the model starts behaving with topic avoidance.

Next, we investigate if the gradual model editing method also impacts the performance in the other domains. Our findings indicate that the specialized behavior is indeed reflected across various domains, even when the AV is extracted for a specific domain. For instance, Table 2 demonstrates that the addition of a medical AV with $\lambda = 0.5$ also enhances the model's expertise in the financial domain. Similarly, we observed that with λ = -1.2 the model exhibits avoidance behavior in both the legal

Lambda	Fin pref. Acc			Leg pref. Acc			General Pref. Acc			
							Safety		Helpfulness	
	Exp	Gen	Avd	Exp	Gen	Avd	Safe	Unsafe	Helpful	Unhelpful
0	.81	19	0	.78	.22	0	.58	.42	.60	.40
0.5	1.0	0	0	1.0	0	0	.58	.42	.66	.34
-0.7	.59	.40	.01	.58	.32	.10	.57	.43	.58	.42
-1.2	.03	.20	.77	.08	.18	.74	.57	.43	.49	.51

Table 2: Out of Domain (special and general) preference accuracy for Medical domain responses. Here, we gradually add the in-domain AV with the base model, and observe the performance for out-of-domain proficiency levels. We find that steering the proficiency levels in one domain also generalizes across other domains.

and financial domains. This pattern is consistent when using other specialized domain vectors as well (see Appendix D).

Effect on General Alignment We also examine whether MEAV for controllable proficiency levels influences the general domain preference (i.e., 'helpfulness' and 'safety'). Notably, we do not observe any regression in the safety domain; however, the model becomes increasingly helpful as λ increases. With the rise in λ, the model provides more detailed and specific guidance, which aligns with human preferences for helpfulness. Conversely, decreasing λ causes the model to avoid answering, which is perceived as unhelpful. Notably, the range of change in general domain preference accuracy is $\pm 11\%$ for helpfulness and $\pm 1\%$ for safety, indicating that MEAV does not lead to significant regression in general domain performance.

6.2 Multi Domain Preference Tuning

We observe distinct behaviors across different domains by adjusting specific configurations. Since, we have three proficiency levels, accuracy higher than 33% and the highest among the three levels can be considered as the "dominant" proficiency level. For example, as shown in Table 3, we find that an AV-based editing coefficient of -1, -1, and 0.6 for the Medical, Financial, and Legal domains, respectively, results in *avoidance* being the dominant behavior in the Medical and Financial domains, with accuracies of 0.46 and 0.42, respectively, and *expertise* being dominant in the Legal domain, with an accuracy of 0.78. Therefore, it indicates multi-level expertise across domains, and we address **RQ2** as well.

There are 27 possible domain–behavior combinations (three domains × three spectrums), and a grid search reveals 22 where the desired behavior is dominant. Joint training achieves near-perfect accuracy but requires 27 separate trainings—nine times more than the three needed for single-domain DPO runs. Each training job takes about 72 hours on an A100 GPU, totaling 1,944 hours for all 27. By contrast, a grid search of 21 coefficient values per domain (9,261 evaluations at roughly 60 seconds each) takes about 155 hours—12 times faster. However, continuous multi-domain tunability remains challenging, as single-domain edits often over-generalize and compromise domain-specific precision.

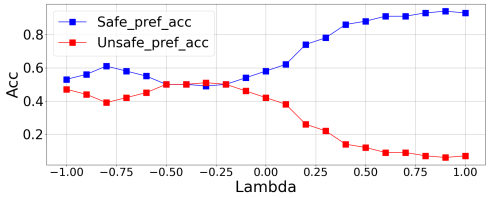


Figure 3: Controlling safety by MEAV

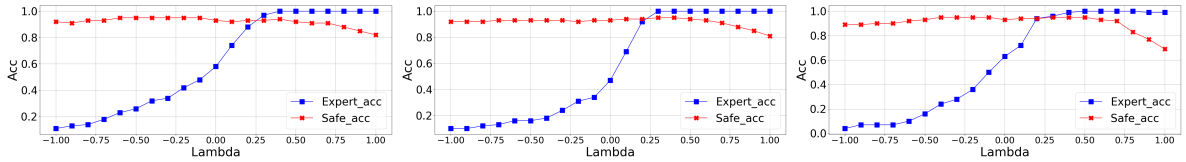
6.3 Can AV be extensible for General Domain?

To explore the generalizability of MEAV across various domains, we focus on the safety alignment as a test case. We start by aligning our base model towards the *safe* dimension by obtaining the safety AV and gradually integrating it with the base model. For the safety alignment, we sample the examples where chosen response is labeled safe, and the rejected response is labeled unsafe (Ji et al., 2024). We compute the pref. acc in the same way described in 5.1, where $R = \{safe, unsafe\}$.

Figure 3 illustrates that the model exhibits mixed safety accuracy initially when $\lambda = 0$ with a safety preference accuracy of 0.53 and an unsafe preference of 0.47. As λ increases, the model progressively aligns more with safety, achieving a safety preference accuracy of 0.93 at λ=1. However, when λ is adjusted negatively, the safety scores

Baseline: Joint training			Ours: MEAV			editing coef
Med	**Fin**	**Leg**	**Med**	**Fin**	**Leg**	
Avd (100%)	Avd (99%)	Exp (98%)	Avd (46%)	Avd (42%)	Exp (78%)	[-1, -1, .6]
Avd (100%)	Exp (91%)	Exp (94%)	Avd (43%)	Exp (44%)	Exp (80%)	[-1, .8, .6]
Avd (100%)	Exp (90%)	Avd (90%)	Avd (57%)	Exp (56%)	Avd (36%)	[-.4, .4, -.8]
Exp (99%)	Avd (100%)	Exp (97%)	Exp (88%)	Avd (44%)	Exp (87%)	[.2, -.8, -.2]

Table 3: Multidomain expertise can be achieved by MEAV. In the baseline joint training approach, we find near-perfect performance, however, we need to perform separate training for each specific configuration. On the contrary, once trained on domain specific expertise, we can perform inference time adjustment and obtain specific configuration to behave in different way in each of the domain.



(a) (b) (c)

Figure 4: Visualizing the transferability of the MEAV process. We observe the effect of proficiency-level-encoded AV integration with a safety-aligned model in the (a) Medical domain (b) Financial Domain (c) Legal Domain proficiency control. For all domains within the range of -1 to 0.7, we donot see any regression of safety, indicating the robustness of MEAV.

become inconsistent and mixed. Notably, even at large negative λ values, beyond -0.25, the model does not become fully "unsafe".

In constructing the response proficiency levels, we intentionally maintain three distinct spectrums. In contrast, the PKU-SafeRLHF dataset does not follow this structure, as it lacks any specific gradation in safety levels.

6.4 Analyzing the Transferability of Alignment Vector

Next, we explore whether AVs derived from a specific alignment objective can be effectively applied to a pre-aligned model. As a case study, we select a safety-aligned version of the base model, to assess the transferability of these alignment vectors. Using a similar approach to single-domain MEAV, we gradually integrate the AVs into our target model, which is safety-aligned.

Figure 4 presents the model's performance as λ is varied. Our findings indicate that when λ is adjusted from -1 to +1, the model's behavior related to safety—its primary control objective—remains relatively stable. For instance, in the medical domain (Figure 4(a)), varying λ within this range results in a minimal change in safety preference accuracy, with a difference of only 0.11 between the lowest and highest accuracy points. In contrast, the accuracy of medical expert response preferences improves significantly, with an increase of 0.81—over seven times greater than the change in safety preference accuracy. Hence, we can conclude that, the AV obtained by our method is trasferable to models aligned on other orthogonally aligned objectives as well, proving the transferability of MEAV.

7 Conclusion

We address inference-time preference alignment tunability through a novel model editing technique called MEAV. We build a synthetic dataset designed to represent three levels of response proficiency across three specialized domains. Our approach enables single-domain preference tunability at inference time without incurring additional costs or resource usage. This allows users to select different response proficiency levels without the need for extra training. Furthermore, our method offers tailored configurations for diverse multidomain behaviors, significantly reducing both training time and resource demands. In future work, we will explore preference tunability in more open-source models like Llama and Qwen (Touvron et al., 2023; Bai et al., 2023). Furthermore, we want to explore the transferability of alignment vectors across different LLMs.

Limitations

Our work has several limitations and areas for future exploration.

- We did not evaluate the correctness of the specialized domain responses. While the authors manually fact-checked a subset of the responses, we do not recommend using these synthetic LLM-generated responses without expert validation. Researchers found a 4.6% rate of hallucinations in Claude-generated response (Vectara, 2025). However, how the hallucinations might impact the special domain responses, is left for future research.

- We used a basic approach (AV) for obtaining alignment vectors, which was simple and effective for our use-case. However, whether the AVs are also capturing noise outside the preference dimension, is not explored in our work. To that end, more advanced techniques like parameter thresholding, zeroing, or SVD-based separation will be explored (Yadav et al., 2024; Gao et al., 2024) in our future work.

- Our method is currently applicable only to LLMs with the same architecture and parameter count. As new models with diverse architectures and varying parameter sizes continue to emerge, this constraint may limit the generalizability of our approach. We aim to extend our methodology to support cross-architecture and cross-parameter adaptation in future.

- We tested our approach only on Mistral-7b, so validation with other open-source LLMs and SLMs is necessary.

- We relied on an extensive grid search for multidomain alignment, which, while more efficient than full retraining, remains computationally intensive. A more optimized or strategic search approach could significantly reduce the parameter search space and further enhance efficiency.

Ethical Implication and Broader Impact

The introduction of MEAV offers a transformative approach to LLM alignment, enabling dynamic, inference-time preference adjustments while significantly reducing computational costs. This flexibility allows LLMs to be more adaptable across different speciality domains—such as medical, legal, and financial—without the need for retraining. However, there are also some concerns with this, and we discuss this below:

- A model originally fine-tuned for safety-aligned behavior could be easily modified at inference time using adversarially crafted AVs to produce harmful, deceptive, or unsafe outputs.

- The expert responses may encode cultural bias in all medical, legal, and financial domains.

- The ability to dynamically adjust model behavior raises concerns about accountability, as users can shift LLM responses in ways that deviate from the ethical constraints originally intended.

References

Anthropic. 2024. Introducing the next generation of claude: The claude 3 family. Accessed: 2024-09-10.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. pages 4447–4455. PMLR.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. 2023. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. *arXiv preprint arXiv:2310.05344*.

Lei Gao, Yue Niu, Tingting Tang, Salman Avestimehr, and Murali Annavaram. 2024. Ethos: Rectifying language models in orthogonal parameter space. *arXiv preprint arXiv:2403.08994*.

Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Jiexin Wang, Huimin Chen, Bowen Sun, Ruobing Xie, Jie Zhou, Yankai Lin, et al. 2024. Controllable preference optimization: Toward controllable multi-objective alignment. *arXiv preprint arXiv:2402.19085*.

James Y Huang, Sailik Sengupta, Daniele Bonadiman, Yi-an Lai, Arshit Gupta, Nikolaos Pappas, Saab Mansour, Katrin Kirchoff, and Dan Roth. 2024. Deal: Decoding-time alignment for large language models. *arXiv preprint arXiv:2402.06147*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2024. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Lei Li, Yongfeng Zhang, and Li Chen. 2023. Prompt distillation for efficient llm-based recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1348–1357.

Tianlin Liu, Shangmin Guo, Leonardo Bianco, Daniele Calandriello, Quentin Berthet, Felipe Llinares, Jessica Hoffmann, Lucas Dixon, Michal Valko, and Mathieu Blondel. 2024. Decoding-time realignment of language models. *arXiv preprint arXiv:2402.02992*.

Bertalan Meskó. 2023. Prompt engineering as an important emerging skill for medical professionals: tutorial. *Journal of medical Internet research*, 25:e50638.

Jonas Oppenlaender, Rhema Linder, and Johanna Silvennoinen. 2023. Prompting ai art: An investigation into the creative skill of prompt engineering. *arXiv preprint arXiv:2303.13534*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. 2024. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vectara. 2025. Hallucination evaluation leaderboard. Hugging Face Spaces. Accessed: 2025-02-15.

Pengyu Wang, Dong Zhang, Linyang Li, Chenkun Tan, Xinghao Wang, Ke Ren, Botian Jiang, and Xipeng Qiu. 2024. Inferaligner: Inference-time alignment for harmlessness through cross-model guidance. *arXiv preprint arXiv:2401.11206*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2024. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

A Data generation and Annotation details

Table 4 shows the breakdown of the total amount of data collected.

Table 5 shows the annotation accuracy for the human volunteers.

B Synthetic Data Generation: How did we arrive at the reported numbers of generated data?

We evaluated the validity of persona-query pairs by manually reviewing a sample of 50 entries. Our analysis confirmed that Claude-3-sonnet reliably adhered to the instructions outlined in our prompt. To determine the dataset size, we initially generated 15,000 queries across all domains, as described in Section 4.1.

During preprocessing, we identified and removed 1–3% of the queries and responses that were truncated due to timeout or quota limit issues. Additionally, queries in non-English languages were excluded to ensure linguistic uniformity. We further filtered for completeness, retaining only those entries that contained a full set of persona-query responses across the three proficiency levels. Consequently, a small number of incomplete entries were discarded.

Next, we examined the dataset for query repetition. Although some personas were repetitive, we found no significant repetition in the queries themselves. Following this rigorous cleaning and filtering process, we finalized the dataset with the following distributions: 13,000 personas for the medical domain, 12,374 personas for the financial domain, and 12,867 personas for the legal domain.

C Persona and Query Generation Prompts

The following prompts are used to generate Personas and Queries:

C.1 CreatePersona

In the CreatePersona method, we generate a random root persona, and from that other persona are generated. Here is the prompt used:

Prompt: Based on the given persona, generate 5 persona, that can be closely or remotely related to the given persona...generate a one-paragraph financial query in first person...

Given Persona: A patient coordinator with excellent interpersonal skills and a knack for digital scheduling systems

Claude-3-Sonnet generated query:

"persona1": "A recent college graduate with student loans and a entry-level job"

"query1": "Hello, I'm Alex, a 22-year-old who just graduated ...marketing position at a local firm, earning $45,000 per year. However, I have accumulated $32,000 in student loans ...financial situation and long-term objectives?"

"persona2": "A single mother juggling multiple part-time jobs and struggling to make ends meet"

"query2": "My name is Emily, and I'm a 32-year-old single mom working two part-time jobs to support my 6-year-old son....have accumulated over $15,000 in credit card debt...increase my income or reduce expenses?"

C.2 PersonaHub

We generate queries directly from a given persona:

Prompt: Based on the persona described below, generate a one-paragraph medical query in first person, that the person fitting the persona can ask to an online medical/health portal. Make sure the query is clear and very specific with nitty-gritty details like names, numbers etc, but brief. It should also include relevant context, concerns, and other details to help the advisor or expert answer properly.

Persona: A retired coach known for their strategic approach to training and injury prevention

Claude-3-Sonnet generated query: As a retired coach ...I have a concerning issue that requires professional medical guidance. Over the past few weeks, I've been experiencing persistent discomfort and stiffness in my left knee, particularly ... to regain full mobility and continue enjoying an active retirement without compromising my long-term joint health.

D Out-of-domain performance

We report the Out-of-domain performance for all three domains, i.e., medical, financial and legal.

E Experimental details

- We used python3.11
- All experiments are done using pytorch 2.0

Domain	Method of Curation		Total
	PersonaHub	CreatePersona	
Medical	5904	6096	13,000
Financial	6909	5465	12,374
Legal	5952	6915	12,867
Total curated data			38,241

Table 4: Full curated data amount

GT	Prediction		
	Exp	Gen	Avd
Exp	82.96%	17.04%	0%
Gen	17.04%	81.70%	1.26%
Avd	0%	1.26%	98.73%

Table 5: Average annotation accuracy for three annotators

- Transformers version 4.43.0 was used

F How Annotators were recruited

We asked our authors (except for the first author) to volunteer as annotators. Each annotators were given detailed instructions with definitions of each response proficiency levels. The task was to select the responses for their respective levels.

F.1 How instructions are provided to the Annotators

Human Evaluation of Generated Text

We have provided you with the following specialized domain query-response pairs. Your tasks are as follows:

Task: Rank the generated responses in the expertise level according to the level definitions.

We generated three levels of responses—**Avoidance**, **Generic**, and **Expert**—for each query, by providing instructions based on the definitions below.

Definitions:

(Level 1) Avoidance: The response completely avoids providing any advice, information, or suggestions. It does not answer the question directly or indirectly. Instead, it politely informs to be unable to assist with the inquiries and recommends consulting an expert professional. Furthermore, the responder refrains from engaging in any discussion or providing any related resources or opinions regarding the issues specific to the domain in question.

(Level 2) Generic: The response represents general, non-specific information that can help guide the user but does not offer detailed advice. It offers general knowledge that is commonly known and easily accessible. The response is informative yet cautious. It emphasizes the importance of consulting an expert professional for accurate information and help and encourages users to seek advice from a qualified professional or institution for specific concerns.

(Level 3) Expert: The response is detailed, with expert-level advice and information. It thoroughly assesses the situations or context described and offers precise explanations and guidance tailored to the specific situation. The response reflects the depth and accuracy expected from an expert professional, and also the advice is not overly generic. Instead, it is comprehensive and nuanced, addressing the user's specific circumstances. Finally, it offers clear, evidence-based recommendations and ensures the guidance is actionable and comprehensive.

Instruction: You will be given three responses for each query. You need to provide the ranking of each response separated by commas. For example, if you think Response 1 is Generic (level 2), Response 2 is Expert (level 3), and Response 3 is Avoidance (level 1), you should only answer: **2,3,1**.

You can also add a note if you want to notify us of something.

Lambda	Med pref. acc			Leg pref. acc			Gen pref. acc			
							Safety		Helpfulness	
	Exp	Gen	Avd	Exp	Gen	Avd	Safe	Unsafe	Helpful	Unhelpful
0	.75	.25	0	.78	.22	0	.58	.42	.60	.40
.30	.97	.02	.01	.98	.02	0	.57	.43	.59	.41
-.40	.61	.37	.02	.57	.35	.08	.59	.41	.57	.43
-1.4	.18	.40	.42	.19	.52	.29	.55	.45	.51	.49

(b) Out of Domain (special and general) preference accuracy for Financial domain responses

Lambda	Med pref. acc			Fin pref. acc			Gen pref. acc			
							Safety		Helpfulness	
	Exp	Gen	Avd	Exp	Gen	Avd	Safe	Unsafe	Helpful	Unhelpful
0	.75	.25	0	.81	.19	0	.58	.42	.60	.40
.30	1.0	0	0	1.0	0	0	.53	.47	.59	.41
-.70	.30	.57	.13	.32	.56	.12	.56	.44	.53	.47
-1.4	.20	.58	.22	.13	.50	.37	.49	.51	.51	.49

(c) Out of Domain (special and general) preference accuracy for Legal domain responses

You will be provided with a spreadsheet with all these columns.