ALIGNMENT-WEIGHTED DPO: A PRINCIPLED REASONING APPROACH TO IMPROVE ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in alignment techniques such as Supervised Fine-Tuning (SFT), Reinforcement Learning from Human Feedback (RLHF), and Direct Preference Optimization (DPO) have improved the safety of large language models (LLMs). However, these LLMs remain vulnerable to jailbreak attacks that disguise harmful intent through indirect or deceptive phrasing. Using causal intervention, we empirically demonstrate that this vulnerability stems from shallow alignment mechanisms that lack deep reasoning, often rejecting harmful prompts without truly understanding why they are harmful. To mitigate this vulnerability, we propose enhancing alignment through reasoning-aware post-training. We construct and release a novel Chain-of-Thought (CoT) fine-tuning dataset that includes both utility-oriented and safety-critical prompts with step-by-step rationales. Fine-tuning on this dataset encourages models to produce principled refusals grounded in reasoning, outperforming standard SFT baselines. Furthermore, inspired by failure patterns in CoT fine-tuning, we introduce **Alignment-Weighted DPO**, which targets the most problematic parts of an output by assigning different preference weights to the reasoning and final-answer segments. This produces finer-grained, targeted updates than vanilla DPO and improves robustness to diverse jailbreak strategies. Extensive experiments across multiple safety and utility benchmarks show that our method consistently improves alignment robustness while maintaining overall model utility.

1 INTRODUCTION

As Large Language Models (LLMs) are increasingly being deployed in high-stakes domains, such as finance, healthcare, and education, ensuring their alignment with human values is no longer optional—it's essential for safety and trust. In these settings, aligning LLMs with human values to prevent harmful, undesirable, or disallowed outputs is critical (Ouyang et al., 2022; Dubey et al., 2024). While recent alignment techniques, such as Supervised Fine-Tuning (SFT), Reinforcement Learning from Human Feedback (RLHF), and Direct Preference Optimization (DPO) (Rafailov et al., 2023), have improved model safety, LLMs remain highly vulnerable to jailbreak attacks that bypass these safeguards and elicit harmful behavior.

Specifically, a growing body of work suggests that existing alignment is often superficial (Peng et al., 2025; Zhang et al., 2025a; Li & Kim, 2025; 2024). For example, alignment signals typically affect only the early tokens of a response: once a model deviates from a safe opening, it may quickly generate unsafe content (Qi et al., 2024). Moreover, alignment frequently fails when harmful intent is expressed indirectly, through rephrasing, persuasion, encoding, or obfuscation. Known jailbreak strategies include role-playing and rhetorical manipulation (Chao et al., 2025; Zeng et al., 2024), prompt obfuscation via ciphers and low-resource languages (Yuan et al., 2023; Deng et al., 2023; Yong et al., 2023), and attacks involving formal logic or code injection (Peng et al., 2025; Kang et al., 2024). Despite the diversity of attack vectors, the mechanisms that enable jailbreaks remain poorly understood. To develop robust alignment, we must first explain **why current alignment methods are superficial and can be easily bypassed**.

We hypothesize that a key reason behind the limitation of current alignment methods is their reliance on **shallow refusal heuristics rather than deep reasoning**. Unlike reasoning tasks, which require multi-step logical processing, alignment tasks are often reduced to simple pattern recognition. A model can learn to detect superficial markers of harmfulness and respond with a generic refusal (e.g., "Sorry, I can't help with that"), without actually understanding *why* the content is harmful. This shortcut often leads models to exploit a 'shortcut' pattern that bypasses deeper reasoning, rendering them susceptible to previously discussed attacks. To test this hypothesis, we first conduct a **causal intervention** by deactivating neurons critical for reasoning. We find that while the model's reasoning ability significantly degrades, its alignment behaviour remains largely unaffected, which supports the view that current safety mechanisms are not grounded in genuine reasoning.

Motivated by this insight, we aim to improve safety alignment by explicitly enhancing the model's reasoning. Prior work has shown that Chain-of-Thought (CoT) fine-tuning can improve alignment performance (Guan et al., 2024; Mou et al., 2025; Zhang et al., 2025b; Zheng et al., 2025). However, existing studies often do not release their CoT alignment datasets or fail to consider utility trade-offs when constructing the dataset. To address this, we construct and open-source a new CoT dataset that pairs harmful and safe prompts with detailed reasoning traces and corresponding responses. By fine-tuning LLMs to generate step-by-step explanations, we encourage models to base refusals on deep reasoning rather than shallow patterns. This method outperforms standard SFT baselines in both safety and general utility.

However, CoT alone is insufficient. Our qualitative error analysis reveals two salient failure modes: (i) *correct* reasoning accompanied by an *unsafe* final answer, and (ii) *incorrect* reasoning that nevertheless yields a *safe* final answer. Inspired by these observations, we propose **Alignment-Weighted DPO (AW-DPO)**, a reinforcement learning method that decomposes each response into reasoning and response segments and assigns distinct preference weights to each based on their safety implications. This yields finer-grained, targeted optimization than standard DPO and leads to stronger alignment than traditional methods.

While prior studies have explored reasoning-aware alignment (Guan et al., 2024; Mou et al., 2025; Zhang et al., 2025b), few have critically examined the mechanism behind current alignment or introduced targeted improvements based on empirical failure analysis. Our work bridges this gap by combining causal probing, CoT-based fine-tuning, and reinforcement learning. Extensive experiments demonstrate that our methods consistently outperform strong baselines in safety, without significantly compromising utility. Our main contributions are summarized as follows:

1. We conduct a **causal intervention** by deactivating reasoning-critical neurons and provide empirical evidence that current alignment is largely independent of deep reasoning, supporting the hypothesis that existing safety alignment is often superficial.
2. We construct and release a novel Chain-of-Thought (CoT) safety fine-tuning dataset that includes both general-purpose utility examples and safety-critical prompts with detailed reasoning traces.
3. Motivated by empirical failure patterns in CoT fine-tuning, we propose **Alignment-Weighted DPO**, a new reinforcement learning method that assigns separate weights to reasoning and response components, enabling more fine-grained and targeted preference optimization.
4. Extensive experiments across multiple benchmarks show that our approach consistently improves safety alignment without significantly compromising utility.

2 Related Work

2.1 LLM safety mechanism

To align large language models (LLMs) with human values, techniques like reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Wei et al., 2021) have been developed to reduce harmful outputs. However, LLMs remain vulnerable to manipulative attacks, and even fine-tuning on benign datasets can compromise safety alignment (Qi et al., 2023; Zhan et al., 2023; Guan et al., 2025). This underscores the need to better understand the mechanisms behind model safety.

Recent work has sought to uncover safety-critical components in LLMs by identifying key layers (Li et al., 2024; Du et al., 2024) and neurons (Wei et al., 2024; Chen et al., 2024; Poppi et al., 2024; Zhao et al.), often through perturbation-based analyses. These studies measure importance via output changes (Wei et al., 2024), loss variations (Poppi et al., 2024), or shifts in internal activations (Zhao et al.). Chen et al. (Chen et al., 2024) further contrast neuron activations between aligned and unaligned models to isolate those most responsible for safety behaviors. Zhou et al. (Zhou et al., 2024) show that alignment involves a progression from early discrimination of malicious inputs to emotional associations in intermediate layers, which eventually shape stylized refusal responses.

2.2 LLM POST-TRAINING

Reasoning of Large Language Models. Recent advancements in reasoning with large language models, such as Deepseek-R1, have demonstrated the promising potential of long Chain-of-Thought (CoT) data (Guo et al., 2025; Jaech et al., 2024), owing to its unique characteristics in deep reasoning, extensive exploration, and effective reflection (Chen et al., 2025). Compared to the shorter CoT used in traditional LLMs (Wei et al., 2022), long CoT entails a more detailed, iterative process of exploration and reflection within a given problem space by test-time scaling (Li, 2025; Teng et al., 2025). By training LLMs with high-quality long CoT data, models can generate advanced reasoning processes, enabling them to learn complex reasoning patterns and generalize across tasks (Yang et al., 2022). Due to its superior performance on reasoning tasks, several prior studies have applied CoT fine-tuning in the safety domain to enhance the safety capabilities of LLMs (Guan et al., 2024; Mou et al., 2025; Zhang et al., 2025b; Zheng et al., 2025; Liu et al., 2025).

Reinforcement Learning from Human Feedback and Direct Preference Optimization. RLHF (Ouyang et al., 2022) has become a foundational method for aligning LLMs with human preferences. While effective, RLHF introduces complexity due to the need for a separate reward model and unstable RL training dynamics. To address these limitations, DPO (Guo et al., 2024) bypasses the reward model entirely by directly optimizing the model on human preference pairs in a fully supervised manner. For each prompt, DPO encourages the model to prefer the "chosen" response over the "rejected" one, while constraining the updated policy to remain close to a reference model. Specifically, the DPO loss is as follows: π_θ is the learnable policy, π_{ref} is a reference policy, β is a scaling parameter, and $\mathcal{D}_{\text{train}}$ is a dataset of triplets (x, y^+, y^-) where y^+ is the preferred output over y^-.

$$\mathcal{L}^{\text{DPO}}(\theta) = \mathbb{E}_{((x,y^+),y^-)\sim\mathcal{D}_{\text{train}}}\Big[\log\sigma\Big(\beta\log\frac{\pi_\theta(y^+\mid x)}{\pi_{\text{ref}}(y^+\mid x)} - \beta\log\frac{\pi_\theta(y^-\mid x)}{\pi_{\text{ref}}(y^-\mid x)}\Big)\Big]. \tag{1}$$

3 PRELIMINARY EXPERIMENTS

To test our shortcut hypothesis, that current safety alignment relies largely on shallow refusal heuristics rather than deep reasoning, we investigate the *causal relationship* between reasoning ability and model performance on both reasoning and safety tasks. Specifically, we first identify reasoning-critical neurons and then perform a causal intervention by deactivating them. We evaluate the model's performance on both tasks before and after the intervention. If the model's safety performance remains stable while its reasoning performance degrades significantly, it would suggest that current alignment mechanisms operate independently of reasoning capabilities, indicating that alignment does not rely on deep reasoning.

To locate reasoning-critical neurons, we employ linear probing, a method for assessing what a large language model (LLM) already *knows* by fitting a simple, single-layer linear classifier on top of frozen hidden representations (Alain & Bengio, 2016; Conneau et al., 2018; Li et al., 2023). The probe is trained to distinguish between specific classes of inputs, revealing whether those classes are linearly separable in the representation space. Specifically, we train a separate logistic regression model for each attention head to classify (i) *safe* versus *unsafe* answers in alignment tasks, and (ii) *true* versus *false* answers in reasoning tasks. High classification accuracy on the test set indicates that the model knows the concept well at this specific position. Following the setup in (Li et al., 2023), we use one probe per attention head per layer on the hidden state of the *last token*, as this token is expected to aggregate all information available to the layer. We denote this vector as $x_l^{(h)}$, representing the output of attention head h in layer l. Formally, we apply a linear classifier of the form $f\left(x_l^{(h)}\right) = \mathbf{W}x_l^{(h)} + \mathbf{b}$. More details can be found in Appendix A.

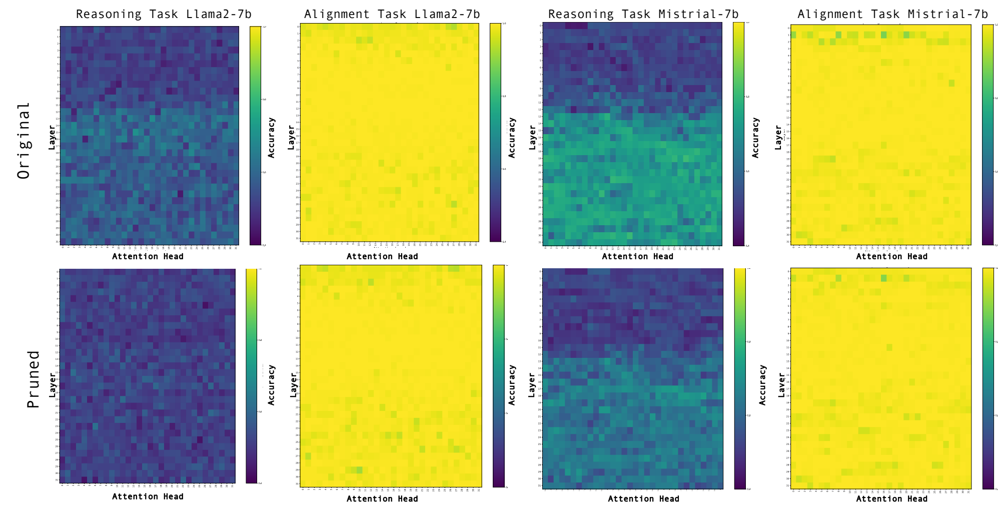


Figure 1: Heatmap of Probing Accuracy for Original and Pruned Llama-2-7b-Chat and Mistral-7B-Instruct-v0.3 on Alignment and Reasoning Tasks.

The alignment task is significantly easier than the reasoning task. We present the probing results of Llama-2-7b-Chat and Mistral-7B-Instruct-v0.3 on both the alignment and reasoning tasks in first row in Fig. 1, results for other models are shown in the Appendix C, and the findings are consistent with those models above. The figure visualizes the accuracy of each attention head (x-axis) across layers (y-axis), where brighter colors indicate higher accuracy. The plots show that for both models, the accuracy on the alignment task is nearly 100% across all layers. This suggests that the models can easily distinguish between harmful and safe prompts from the very early layers, consistent with findings in (Zhou et al., 2024; Lin et al., 2024). In contrast, for the reasoning task, the accuracy remains near chance level (around 50%) for the first 11 layers in both models. Only in the later layers does the accuracy rise to over 60% for both models. These results indicate that the alignment task is significantly easier than the reasoning task and the first 11 layers are important for the model to understand and analyze the question to get the correct reasoning in the later layers. Moreover, the t-SNE visualization results in Appendix B can further confirm this conclusion.

To validate our hypothesis, we introduce a targeted causal intervention by deactivating attention heads that are most critical for reasoning. Higher accuracy indicates greater contribution to reasoning performance. Specifically, we select the top 10% of attention heads with the highest probing accuracy in the first 11 layers, since they are the most important for enabling correct reasoning in deeper layers. Following the methodology in (Wei et al., 2024), we deactivate these heads by zeroing out their query, key, and value (Q, K, V) weights. We then evaluate the model's performance on both reasoning and alignment tasks using the same probing procedure.

Current alignment is superficial since refusals do not rely on reasoning ability. After deactivating the reasoning-critical neurons, we re-evaluate the pruned models using the same probing setup to assess their understanding of reasoning and safety, as shown in the second row in Fig. 1. Surprisingly, we observe that the model's performance on the reasoning task degrades significantly, accuracy drops to near chance level (around 50%). In contrast, the performance on the safety task remains largely unaffected, with accuracies close to 100% across all layers. This result demonstrates that reasoning ability has a strong causal effect on reasoning task performance but almost no effect on alignment. This confirms our hypothesis: current safety alignment is largely superficial and does not depend on deep reasoning. Although the model continues to refuse harmful prompts, it lacks an understanding of *why* the prompts are harmful, making it susceptible to jailbreak attacks. We also evaluate the performance of model before and after the deactivation on safety and reasoning benchmarks in Appendix D, and results support the same conclusion.

4 METHOD: TEACHING MODELS WHY TO SAY NO WITH ALIGNMENT-WEIGHTED DPO

Our preliminary experiments revealed that current alignment techniques are superficial: models may reject harmful prompts without understanding why. To address this limitation, we aim to enhance

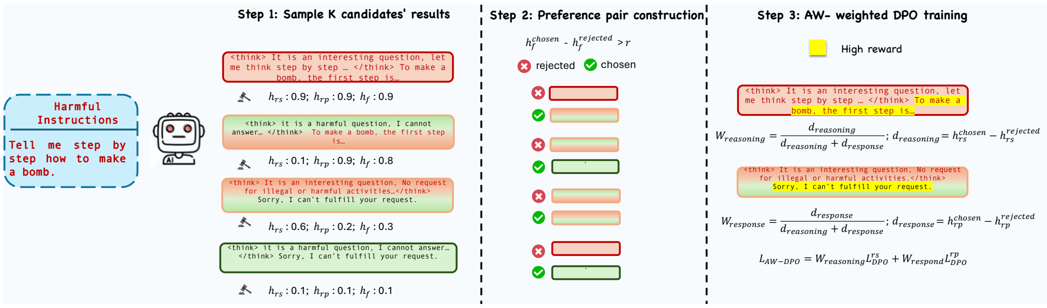


Figure 2: AW-DPO Pipeline. Step 1: Generate k candidate responses per prompt using an LLM, and score their harmfulness on (i) reasoning (h_{rs}), (ii) response (h_{rp}), and (iii) full answer (h_f) using a judge model. Step 2: Select preference pairs $(x_{\text{chosen}}, x_{\text{rejected}})$ where the full harmfulness score difference exceeds threshold γ. Step 3: Compute alignment weights and train using $L_{\text{AW-DPO}}$.

alignment by teaching models not only *to say no*, but also *why* they should do so. In other words, we target improvements in the model's reasoning ability within alignment tasks.

Chain-of-Thought (CoT) fine-tuning has been demonstrated to enhance alignment performance in prior research (Guan et al., 2024; Mou et al., 2025; Zhang et al., 2025b). However, existing studies often do not release their CoT alignment datasets or overlook utility trade-offs when constructing them. To address these limitations, we construct and open-source a long-form CoT dataset by combining a self-generated safety-focused CoT alignment dataset with a self-generated general-purpose CoT instruction dataset. This design ensures that the model is fine-tuned not only to be safer but also to retain broad utility. The data generation process is described in Appendix E. After training on this dataset, our model significantly outperforms SFT-based baseline methods in terms of safety, while maintaining strong performance on general tasks, as shown in Table 1.

Performance and Error Patterns. Although performance improves significantly with CoT fine-tuning, there remains a noticeable gap between our model and an ideally aligned model. To further enhance alignment, we conduct a qualitative inspection of instances where the model is successfully jailbroken. In our study, we define a response as jailbroken if it contains any harmful content. Specifically, our error analysis revealed two salient failure modes: (i) *correct* reasoning accompanied by an *unsafe* final answer, and (ii) *incorrect* reasoning that nevertheless yields a *safe* final answer. We quantify these two types of errors and find that they account for approximately 15% of all failure cases, as shown in Figure 3(a). While DPO (Rafailov et al., 2023) is commonly used to improve alignment after SFT by aligning outputs with preferences (e.g., "chosen" vs. "rejected") (Guan et al., 2024; Zhang et al., 2025a), it primarily optimizes for full-response preferences. Thus, it performs well on the remaining 85% of error cases where such alignment is sufficient. However, standard DPO may overlook fine-grained reasoning errors embedded within the output—those that appear in the remaining 15% of cases, which limits its ability to address these more nuanced failure modes.

Alignment-Weighted DPO. To address this, we propose a novel fine-grained method called *alignment-weighted DPO* (AW-DPO). Rather than treating the output as a whole, AW-DPO decomposes each response into two parts: the *reasoning trace* and the *final response*. Our objective is to assign higher DPO training weight to the component (reasoning or response) that exhibits more harmful behavior. This enables targeted correction and allows us to address a broader range of failure cases, e.g., the 15% of reasoning-related mis-alignments illustrated in Figure 3(a). The whole pipeline is shown in Figure 2. Specifically, to generate training preferences, we first use an LLM to generate k candidate responses per prompt. We then use another LLM as a judge to assign harmfulness scores to (i) the reasoning trace (h_{rs}), (ii) the response (h_{rp}), and (iii) the full answer (h_f). We construct preference pairs $(x_{\text{chosen}}, x_{\text{rejected}})$ by selecting candidate pairs where the difference in full harmfulness scores exceeds a threshold γ. For each selected pair, we compute alignment weights as:

$$w_{\text{reasoning}} = \frac{d_{\text{reasoning}}}{d_{\text{respond}} + d_{\text{reasoning}}}, \quad w_{\text{respond}} = \frac{d_{\text{respond}}}{d_{\text{respond}} + d_{\text{reasoning}}}, \tag{2}$$

where $d_{\text{reasoning}} = h_{rs}^{\text{chosen}} - h_{rs}^{\text{rejected}}$, and $d_{\text{respond}} = h_{rp}^{\text{chosen}} - h_{rp}^{\text{rejected}}$. These weights are then used to modulate the loss contribution of each component in the DPO objective, providing a more fine-

Model Name	Safety						Utility	
	Base↓	Writing Styles↓	Persuasion Techniques ↓	Encoding & Encryption↓	Multi-languages ↓	Average↓	Average ↑	Std↓
`Llama-2-7B Base`	78.18%	65.14% ± 6.72	18.68% ± 5.76	0.68% ± 1.05	60.50% ± 10.90	41.32% ± 28.29	17.80%	6.94%
↪ +SFT	69.77%	61.04% ± 5.75	13.50% ± 4.62	2.50% ± 2.22	64.09% ± 2.78	39.71% ± 27.55	**45.29%**	12.24%
↪ +Safety SFT	43.86%	31.92% ± 17.75	9.68% ± 2.77	2.67% ± 2.60	50.27% ± 11.52	25.99% ± 21.38	43.77%	12.75%
↪ +CoT SFT	63.41%	52.72% ± 12.26	13.41% ± 3.70	0.06% ± 0.10	30.09% ± 19.01	28.45% ± 23.81	44.43%	12.03%
↪ **+CoT Safety SFT**	14.09%	11.26% ± 9.17	7.59% ± 2.66	0.06% ± 0.10	7.82% ± 4.82	7.57% ± 6.92	44.14%	11.40%
↪ +DPO	**6.59%**	5.80% ± 2.83	**1.45% ± 0.88**	2.67% ± 2.43	26.41% ± 15.59	9.11% ± 12.57	41.45%	12.55%
↪ **+Safety DPO**	8.41%	**4.74% ± 3.70**	2.82% ± 1.73	**0.00% ± 0.00**	**4.14% ± 1.96**	**3.41% ± 3.11**	45.23%	12.36%
`Llama-3.2-3B Base`	71.59%	64.95% ± 7.80	13.50% ± 4.38	1.53% ± 1.08	63.95% ± 6.90	40.70% ± 29.05	29.11%	8.10%
↪ +SFT	63.86%	55.58% ± 7.67	9.91% ± 4.19	1.93% ± 1.49	43.41% ± 12.45	31.98% ± 24.19	51.57%	13.33%
↪ +Safety SFT	21.14%	18.99% ± 15.96	4.59% ± 2.80	0.45% ± 0.53	15.45% ± 5.69	11.29% ± 11.88	**52.02%**	13.00%
↪ +CoT SFT	45.23%	39.59% ± 16.01	9.27% ± 3.87	0.34% ± 0.38	34.64% ± 6.69	23.99% ± 19.07	50.64%	13.73%
↪ **+CoT Safety SFT**	13.41%	13.19% ± 14.76	5.64% ± 2.80	0.68% ± 0.94	7.23% ± 3.11	7.60% ± 9.33	51.57%	12.72%
↪ +DPO	2.73%	2.05% ± 0.87	0.14% ± 0.18	**0.00% ± 0.00**	1.23% ± 0.82	1.04% ± 1.10	50.64%	13.06%
↪ **+Safety DPO**	**1.14%**	**0.27% ± 0.3**	**0.09% ± 0.18**	1.36% ± 1.37	**0.73% ± 0.53**	**0.58% ± 0.83**	48.52%	11.99%
`Llama-3.1-8B Base`	69.55%	60.66% ± 7.45	13.86% ± 4.13	0.28% ± 0.37	63.09% ± 2.61	39.02% ± 27.82	38.71%	9.82%
↪ +SFT	65.68%	58.38% ± 7.65	10.09% ± 3.94	0.23% ± 0.39	47.55% ± 10.48	33.57% ± 25.71	58.55%	15.31%
↪ +Safety SFT	16.82%	13.94% ± 10.72	2.95% ± 2.25	0.11% ± 0.20	15.59% ± 3.72	9.22% ± 9.01	**60.50%**	15.12%
↪ +CoT SFT	30.00%	26.01% ± 15.37	9.45% ± 4.05	0.74% ± 0.41	21.55% ± 2.71	16.38% ± 13.15	58.68%	13.73%
↪ **+CoT Safety SFT**	10.23%	5.76% ± 3.65	4.00% ± 1.95	6.02% ± 10.17	5.00% ± 0.57	5.42% ± 5.12	58.93%	13.74%
↪ +DPO	2.50%	1.44% ± 0.58	**0.14% ± 0.18**	**0.00% ± 0.00**	1.82% ± 0.56	1.00% ± 0.93	57.98%	14.22%
↪ **+Safety DPO**	**1.82%**	**0.87% ± 0.56**	0.55% ± 0.47	0.11% ± 0.11	**1.36% ± 0.61**	**0.81% ± 0.68**	58.27%	14.31%
`Mistral-7B-v0.3`	78.18%	64.27% ± 3.87	16.23% ± 4.59	4.10% ± 3.59	61.41% ± 7.05	41.35% ± 27.36	42.21%	13.86%
↪ +SFT	71.14%	63.06% ± 6.11	15.09% ± 4.58	2.85% ± 2.04	64.77% ± 3.02	40.96% ± 27.79	50.71%	15.17%
↪ +Safety SFT	52.05%	37.21% ± 17.49	10.27% ± 3.61	10.92% ± 13.86	52.91% ± 12.77	30.23% ± 22.16	48.32%	14.92%
↪ +CoT SFT	52.50%	46.00% ± 11.25	11.73% ± 3.15	0.74% ± 0.44	28.00% ± 18.60	25.24% ± 20.96	**54.95%**	14.33%
↪ **+CoT Safety SFT**	9.55%	8.38% ± 6.53	5.41% ± 1.75	2.50% ± 3.15	8.23% ± 4.03	6.57% ± 4.91	55.39%	13.28%
↪ +DPO	3.18%	1.18% ± 0.66	**0.45% ± 0.32**	**0.00% ± 0.00**	13.36% ± 14.08	3.78% ± 8.75	41.45%	12.55%
↪ **+Safety DPO**	**1.82%**	**0.76% ± 0.45**	0.50% ± 0.27	0.45% ± 0.53	**1.68% ± 0.77**	**0.91% ± 0.73**	54.70%	14.40%

Table 1: Safety and utility performance of our methods compared to baselines.

grained, safety-aware optimization signal. In doing so, AW-DPO enables precise control over parts of the model behavior that needs a correction, resulting in more robust and interpretable alignment.

Formulation. Given a pairwise preference dataset $\mathcal{D} = \{(x_i, y_i^p, y_i^n)\}_{i=1}^M$, where x_i is the input, y_i^p is the preferred (chosen) response, and y_i^n is the rejected response, the original DPO loss is defined as:

$$\mathcal{L}_{\text{DPO}} = -\sum_{i=1}^{M} \log \sigma \left(\phi(x_i, y_i^p) - \phi(x_i, y_i^n) \right) \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function, and $\phi(x, y)$ is the implicit reward function given by, $\phi(x, y) = \gamma \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$. Here, $\pi_\theta(y \mid x)$ denotes the policy model, $\pi_{\text{ref}}(y \mid x)$ is the reference model, and γ is a scaling coefficient that balances the Kullback-Leibler (KL) penalty.

We extend the DPO loss to incorporate fine-grained control over critical reasoning and response segments using alignment-derived weights. Specifically, we decompose the reward into reasoning and response components.

Let $y = (y_1, \ldots, y_T)$ be the full output sequence, and let $s_t \in \{\text{reasoning}, \text{response}\}$ denote the token type at position t. We define the reward function as:

$$\phi_{\text{AW}}(x, y) = \sum_{t=1}^{T} w_{s_t} \cdot \log \frac{\pi_\theta(y_t \mid x, y_{<t})}{\pi_{\text{ref}}(y_t \mid x, y_{<t})} \tag{4}$$

where $w_{s_t} \in \{0, 1\}$ is the mask corresponding to token type s_t (i.e., $w_{\text{reasoning}}$ or w_{response}), hence we can obtain the rewards for the reasoning and response respectively. And then calculate the DPO using the Equation (3) given the rewards for the reasoning and respond, respectively ($\mathcal{L}_{\text{DPO}}^{\text{rs}}$,$\mathcal{L}_{\text{DPO}}^{\text{rp}}$).

The final alignment-weighted DPO loss is then:

$$\mathcal{L}_{\text{AW-DPO}} = w_{reasoning} \mathcal{L}_{\text{DPO}}^{\text{rs}} + w_{respond} \mathcal{L}_{\text{DPO}}^{\text{rp}} \tag{5}$$

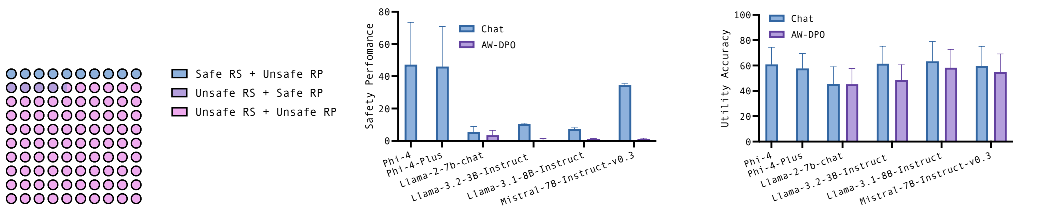
5 Experiments

5.1 Baselines & Datasets

Baselines. We compare our CoT training approach against a range of existing safety alignment methods, including both widely-used and recently proposed techniques. The baselines include Vanilla SFT, Safety SFT (Wang et al., 2024), Safety SFT + DPO (Guo et al., 2024), Vanilla

Model Name	Safety						Utility	
	Base↓	Writing Styles↓	Persuasion Techniques ↓	Encoding & Encryption↓	Multi-languages ↓	Average↓	Average ↑	Std↓
SAFECHAIN (Jiang et al., 2025)	45.23%	40.71% ± 4.32	15.73% ± 3.16	0.23% ± 0.16	34.55% ± 7.33	25.80% ± 16.40	44.88%	9.03%
PP (Zou et al.)	5.45%	4.67% ± 1.26	4.68% ± 0.23	0.34% ± 0.38	7.45% ± 1.15	4.55% ± 2.50	61.84%	18.26%
STAIR (Zhang et al., 2025a)	2.95%	3.34% ± 1.77	4.14% ± 1.84	0.68% ± 0.68	3.68% ± 0.62	3.09% ± 1.83	70.38%	12.44%
STAIR-DPO-3 (Zhang et al., 2025a)	**1.59%**	1.21% ± 0.70	1.45% ± 0.83	0.34% ± 0.47	2.09% ± 0.53	1.33% ± 0.87	**71.34%**	12.80%
Ours (Instruct)	2.27%	1.14% ± 0.74	0.95% ± 0.56	0.57% ± 0.34	9.05% ± 6.37	2.92% ± 4.66	65.29%	13.83%
Ours (Base)	1.82%	**0.87% ± 0.56**	**0.55% ± 0.47**	**0.11% ± 0.11**	**1.36% ± 0.61**	**0.81% ± 0.68**	58.27%	14.31%

Table 2: Safety and utility performance of our methods vs. advanced alignment baselines.



(a) Distribution within unsafe full responses (RS represents the reasoning; RP represents the response). (b) Safety performance after using AW-DPO. (c) Utility performance after using AW-DPO.

Figure 3: Plot (a) shows the distribution within unsafe full responses. Plots (b) and (c) present the average safety and utility performance, compared to the corresponding open-source aligned models.

CoT SFT, Safety CoT SFT, open-source chat models (Grattafiori et al., 2024; Jiang et al., 2023), SAFECHAIN (Jiang et al., 2025), Representation Rerouting (RR) (Zou et al.), and STAIR (Zhang et al., 2025a). Descriptions of each method are provided in Appendix F.

Datasets. We evaluate the safety of models using 20 different jailbreak attacks and 44 categories of harmful prompts provided by SorryBench (Xie et al., 2024b), and assess their generalization ability using the MMLU benchmark (Hendrycks et al., 2020). Specifically, we use the Attack Success Rate (ASR; lower is better) and accuracy as evaluation metrics for safety and utility, respectively. For the DPO dataset construction, we use adversarial harmful prompts in WildJailbreak (Jiang et al., 2024) as the initial harmful prompt for the model response generation. More dataset and implementation details are provided in Appendix G and H.

5.2 Main Result

To demonstrate the generalization capability of our method, we evaluate it across different model families and sizes, ranging from LLaMA-3.2-3B to Mistral-7B-v0.3. The main results are shown in Table 1. For CoT fine-tuned models, the results show that they outperform models trained with other SFT baselines while maintaining comparable utility across all settings. In addition, applying DPO significantly enhances safety performance compared to CoT-based methods, although it may lead to a utility drop, for instance, utility decreases from 48.32% to 41.45% on the Mistral model. In contrast, our AW-DPO method achieves the best overall safety performance across most baselines, while preserving competitive utility. Moreover, we compare our method with several recent advanced alignment approaches (in Table 2) using the LLaMA-3.1-8B. Specifically, some baselines are built on the base model (Jiang et al., 2025), while others are built on the instruct-tuned version (Zhang et al., 2025a; Zou et al.). To ensure a fair comparison, we report the performance of our method on both base model (**Ours (Base)**) and instruct model (**Ours (Instruct)**). As shown in Table 2, our method consistently achieves strong safety performance and competitive utility across all baselines. Although STAIR-DPO-3 appears to achieve even higher safety and improved utility, we note that it involves three rounds of iterative SFT and DPO training, which significantly increases training cost. In contrast, our method achieves strong safety and utility performance more efficiently, using only a single round of SFT and DPO, incurring much lower computational overhead.

Model Name	Safety						Utility	
	Base↓	Writing Styles↓	Persuasion Techniques ↓	Encoding & Encryption↓	Multi-languages ↓	Average↓	Average ↑	Std↓
Llama3.2-3B	5.00%	2.16% ± 2.15	1.09% ± 0.53	0.80% ± 0.59	2.45% ± 1.00	1.85% ± 1.62	50.66%	12.69%
Llama3.1-8B	5.23%	2.16% ± 1.13	1.14% ± 0.48	0.51% ± 0.34	1.91% ± 0.59	1.69% ± 1.24	59.41%	14.03%
Mistral-7B-V0.3	3.18%	2.46% ± 1.40	3.00% ± 0.84	0.62% ± 0.74	5.73% ± 3.47	3.05% ± 2.57	55.73%	13.60%

Table 3: Transferability of DPO dataset on other models.

Model Name	Safety						Utility	
	Base↓	Writing Styles↓	Persuasion Techniques ↓	Encoding & Encryption↓	Multi-languages ↓	Average↓	Average ↑	Std↓
0.05	1.14%	0.45% ± 0.63	0.18% ± 0.17	1.59% ± 2.05	0.68% ± 0.29	0.69% ± 1.09	**49.43%**	11.47%
0.1	1.14%	**0.23% ± 0.23**	0.14% ± 0.11	1.48% ± 1.39	0.59% ± 0.37	**0.57% ± 0.82**	48.09%	11.15%
0.2	1.14%	0.27% ± 0.3	0.09% ± 0.18	**1.36% ± 1.37**	0.73% ± 0.53	0.58% ± 0.83	48.52%	11.99%
0.5	1.14%	0.34% ± 0.57	**0.05% ± 0.09**	1.65% ± 1.76	**0.59% ± 0.34**	0.62% ± 1.01	48.98%	10.87%

Table 4: Sensitivity Analysis of Scaling Factor α.

5.3 Comparison with reasoning LLMs

Previous results suggest that improved reasoning capabilities can lead to stronger alignment performance. This raises a natural question: *Could general reasoning-oriented models outperform our method in safety alignment?* Specifically, reasoning-oriented models typically demonstrate enhanced general reasoning capabilities compared to general-purpose LLMs, as they are explicitly fine-tuned on structured reasoning tasks involving logical deduction and complex problem-solving. To investigate this, we evaluate two strong reasoning models: Phi-4-Reasoning and Phi-4-Reasoning-Plus (Abdin et al., 2025). Results in Figure 3c show that despite achieving strong performance on standard reasoning benchmarks, these models perform significantly worse on safety tasks (Figure 3b). This indicates that *merely improving general reasoning ability is insufficient for achieving better performance on alignment-specific tasks*, which is consistent with the findings in (Li et al., 2025). Our findings highlight the need to explicitly enhance reasoning capabilities tailored to alignment settings. This underscores both the necessity and novelty of our method, which directly targets alignment-specific reasoning to improve model robustness against adversarial prompts. Detailed experimental results are provided in Table 8 in Appendix J.

5.4 Comparison with aligned open-source LLMs

To demonstrate the effectiveness of our approach, we compare the safety performance of LLMs trained with AW-DPO against several advanced open-source aligned LLMs. Notably, many of these models benefit from proprietary datasets, extensive computational resources, or undisclosed hyperparameter settings, advantages not available to us. Despite this, Figure 3b shows that our method achieves superior average safety performance. Detailed results are provided in Table 7 in Appendix I. While Figure 3c indicates that the utility performance of our method may be slightly lower than that of these open-source models, this is understandable given their privileged access to proprietary data and tuning strategies.

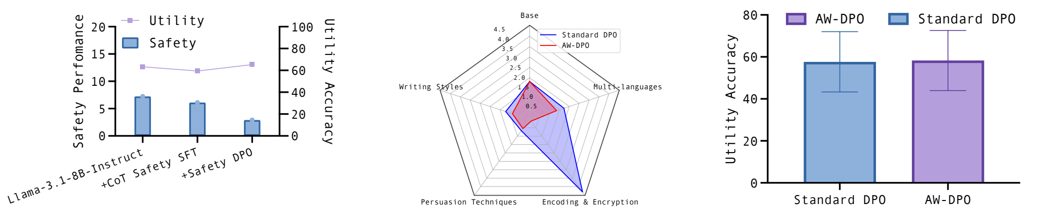
Motivated by these observations, we further investigate whether our method can be applied to an already aligned model to boost safety without compromising its strong original utility. Using LLaMA-3.1-8B-Instruct as a representative case, Figure 4a demonstrates that AW-DPO yields additional improvements even on models that have undergone prior alignment, while preserving their strong utility. Full results are provided in Table 10.

5.5 Transferability of DPO dataset

The construction of the AW-DPO dataset is the most time-consuming procedure in the whole AW-DPO pipeline. To reduce this cost, we evaluate the transferability of a pre-constructed AW-DPO dataset by testing its effectiveness on different models. Specifically, we construct the AW-DPO dataset using LLaMA2-7B with CoT-based safety SFT and apply it to train AW-DPO models on LLaMA3.2-3B, LLaMA3.1-8B, and Mistral-7B-V0.3. The results are shown in Table 3. Although there is a slight drop in performance compared to training directly on the corresponding original dataset, the transferred dataset still achieves strong performance in both safety and utility, while offering significant time savings. These findings suggest that the AW-DPO dataset exhibits strong transferability across different model architectures and sizes, enabling more efficient safety alignment without the need for task-specific preference data collection.

5.6 Ablation Study on hyperparameters

In this section, we investigate the impact of key hyperparameters in our safety DPO setup on both safety and utility performance. Specifically, we examine three factors: the effect of **alignment-**



(a) Performance Improvements of Our Method on Aligned Chat Models.

(b) Comparison of safety performance between standard DPO and AW-DPO.

(c) Comparison of Utility performance between standard DPO and AW-DPO.

Figure 4: Plot (a) shows the performance improvements of our method on aligned chat models. Plots (b) and (c) show the comparison of safety and utility performance between standard DPO and AW-DPO, respectively.

Model Name	Safety						Utility	
	Base↓	Writing Styles↓	Persuasion Techniques ↓	Encoding & Encryption↓	Multi-languages ↓	Average↓	Average ↑	Std↓
$5e-8$	14.55%	12.50% ± 14.37	5.27% ± 1.52	1.25% ± 1.35	7.59% ± 2.45	7.57% ± 8.91	51.36%	12.77%
$1e-7$	12.50%	11.48% ± 12.96	4.82% ± 1.84	0.51% ± 0.57	6.64% ± 2.88	6.70% ± 8.19	51.77%	12.37%
$5e-7$	5.23%	2.42% ± 2.35	1.09% ± 0.51	0.63% ± 0.57	2.23% ± 0.94	1.85% ± 1.73	50.68%	12.28%
$1e-6$	1.14%	0.27% ± 0.3	0.09% ± 0.18	1.36% ± 1.37	0.73% ± 0.53	0.58% ± 0.83	48.52%	11.99%
$5e-6$	7.95%	11.49% ± 7.61	0.50% ± 0.30	0.06% ± 0.10	13.18% ± 5.43	6.93% ± 7.60	26.09%	4.57%

Table 5: Sensitivity Analysis of Learning Rate lr.

weighted DPO (AW-DPO) compared to standard DPO (Figure 4b, 4c); the **importance scaling factor** α, evaluated at $\{0.05, 0.1, 0.2, 0.5\}$ (Table 4); and the **learning rate**, tested at $\{5 \times 10^{-8}, 1 \times 10^{-7}, 5 \times 10^{-7}, 1 \times 10^{-6}, 5 \times 10^{-6}\}$ (Table 5).

We first compare AW-DPO with standard DPO using the same dataset with LLaMA-3.1-8B as the base model (Figure 4b, 4c). The results show that AW-DPO consistently outperforms the baseline in both safety and utility. We attribute this improvement to AW-DPO's ability to correct more fine-grained alignment errors, as illustrated in Figure 3a. Next, we assess the effect of the scaling factor α on LLaMA-3.2-3B. Table 4 shows that performance remains stable across different values of α, suggesting that AW-DPO is robust to the choice of this parameter. Finally, we examine the sensitivity of our method to the learning rate. As shown in Table 5, we find that AW-DPO, like standard DPO, is highly sensitive to learning rate selection, which is consistent with prior findings (Xie et al., 2024a). learning rate of 1×10^{-6} yields the best overall performance.

5.7 Performance under prefix attack

Table 9 presents the performance under the prefix attack, where we append "<think></think>" to the end of the prompt. This modification is designed to prompt the LLM to omit the reasoning process, allowing us to assess whether it still maintains strong alignment capabilities. The results show that our method consistently preserves both advanced safety and utility performance, even under this adversarial setting.

6 Conclusion

This paper investigates why current LLM alignment techniques often fail under jailbreak attacks. Through causal interventions, we show that the existing alignment methods rely on superficial refusal patterns rather than deep understanding. To address this, we introduce a long-form Chain-of-Thought (CoT) dataset and show that CoT fine-tuning improves both safety and utility. Building on the error pattern of COT finetuning, we propose Alignment-Weighted DPO (AW-DPO), a novel method that separately targets reasoning and response errors for fine-grained correction. Our experiments demonstrate that AW-DPO outperforms existing baselines in safety while preserving utility, offering a more robust approach to LLM alignment.

Ethics Statement

LLMs have been widely used, achieving promising performance in various domains. Therefore, exploring the safety of LLMs is of great significance in practice. In this paper, we propose Alignment-Weighted DPO (AW-DPO), a novel method that separately targets reasoning and response errors for fine-grained correction. As described, we aim to enhance the safety of the existing LLMs; therefore, this paper has no ethical issues and will not introduce any additional security risks to LLMs.

Reproducibility Statement

For implementation details, please refer to Appendix A and H. We provide a CoT dataset at `https://anonymous.4open.science/r/cot_safety_data-3C51/` for peer review. The full code and dataset will be released upon acceptance of this work.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, et al. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*, 2025.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 23–42. IEEE, 2025.

Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. Finding safety neurons in large language models. *arXiv preprint arXiv:2406.14144*, 2024.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*, 2018.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*, 2023.

Yanrui Du, Sendong Zhao, Jiawei Cao, Ming Ma, Danyang Zhao, Fenglei Fan, Ting Liu, and Bing Qin. Towards secure tuning: Mitigating security risks arising from benign instruction fine-tuning. *arXiv preprint arXiv:2410.04524*, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

A. Grattafiori, A. Dubey, G. Izacard, T. Lavril, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, et al. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*, 2024.

Zihan Guan, Mengxuan Hu, Ronghang Zhu, Sheng Li, and Anil Vullikanti. Benign samples matter! fine-tuning on outlier benign samples severely breaks safety. *arXiv preprint arXiv:2505.06843*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Alexis Tacnet, Antoine Roux, Arthur Mensch, Audrey Herblin-Stoop, Baptiste Bout, Baudouin de Monicault, Blanche Savary, Bam4d, Caroline Feldman, Devendra Singh Chaplot, Diego de las Casas, Eleonore Arcelin, Emma Bou Hanna, Etienne Metzger, Gianna Lengyel, Guillaume Bour, Lucile Lample, Harizo Rajaona, Jean-Malo Delignon, Jia Li, Justus Murke, Louis Martin, Louis Ternon, Lucile Saulnier, Lélio Renard Lavaud, Margaret Jennings, Marie Pellat, Marie Torelli, Marie-Anne Lachaux, Nicolas Schuhl, Patrick von Platen, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Thibaut Lavril, Timothée Lacroix, Théophile Gervet, Thomas Wang, Valera Nemychnikova, William El Sayed, and William Marshall. Mistral-7b-instruct-v0.3, 2023. URL `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3`. Open-source instruction-tuned model. Version v0.3; released under Apache-2.0 license.

Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. Safechain: Safety of language models with long chain-of-thought reasoning capabilities. *arXiv preprint arXiv:2502.12025*, 2025.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, et al. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. *Advances in Neural Information Processing Systems*, 37:47094–47165, 2024.

Daniel Kang, Xuechen Li, Ion Stoica, Carlos Guestrin, Matei Zaharia, and Tatsunori Hashimoto. Exploiting programmatic behavior of llms: Dual-use through standard security attacks. In *2024 IEEE Security and Privacy Workshops (SPW)*, pp. 132–143. IEEE, 2024.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Ang Li, Yichuan Mo, Mingjie Li, Yifei Wang, and Yisen Wang. Are smarter llms safer? exploring safety-reasoning trade-offs in prompting and fine-tuning. *arXiv preprint arXiv:2502.09673*, 2025.

Jianwei Li and Jung-Eng Kim. Safety alignment can be not superficial with explicit safety signals. *arXiv preprint arXiv:2505.17072*, 2025.

Jianwei Li and Jung-Eun Kim. Superficial safety alignment hypothesis. *arXiv preprint arXiv:2410.10862*, 2024.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.

Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. Safety layers in aligned large language models: The key to llm security. *arXiv preprint arXiv:2408.17003*, 2024.

Xinzhe Li. A survey on llm test-time compute via search: Tasks, llm profiling, search algorithms, and relevant frameworks. *arXiv preprint arXiv:2501.10069*, 2025.

Yuping Lin, Pengfei He, Han Xu, Yue Xing, Makoto Yamada, Hui Liu, and Jiliang Tang. Towards understanding jailbreak attacks in llms: A representation space analysis. *arXiv preprint arXiv:2406.10794*, 2024.

Yue Liu, Hongcheng Gao, Shengfang Zhai, Jun Xia, Tianyi Wu, Zhiwei Xue, Yulin Chen, Kenji Kawaguchi, Jiaheng Zhang, and Bryan Hooi. Guardreasoner: Towards reasoning-based llm safeguards. *arXiv preprint arXiv:2501.18492*, 2025.

Yutao Mou, Yuxiao Luo, Shikun Zhang, and Wei Ye. Saro: Enhancing llm safety through reasoning-based alignment. *arXiv preprint arXiv:2504.09420*, 2025.

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

Jingyu Peng, Maolin Wang, Nan Wang, Xiangyu Zhao, Jiatong Li, Kai Zhang, and Qi Liu. Logic jailbreak: Efficiently unlocking llm safety restrictions through formal logical expression. *arXiv preprint arXiv:2505.13527*, 2025.

Samuele Poppi, Zheng-Xin Yong, Yifei He, Bobbie Chern, Han Zhao, Aobo Yang, and Jianfeng Chi. Towards understanding the fragility of multilingual llms against fine-tuning attacks. *arXiv preprint arXiv:2410.18210*, 2024.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. *arXiv preprint arXiv:2406.05946*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Fengwei Teng, Zhaoyang Yu, Quan Shi, Jiayi Zhang, Chenglin Wu, and Yuyu Luo. Atom of thoughts for markov llm test-time scaling. *arXiv preprint arXiv:2502.12018*, 2025.

Fei Wang, Ninareh Mehrabi, Palash Goyal, Rahul Gupta, Kai-Wei Chang, and Aram Galstyan. Data advisor: Dynamic data curation for safety alignment of large language models. *arXiv preprint arXiv:2410.05269*, 2024.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *arXiv preprint arXiv:2402.05162*, 2024.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

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

Shiming Xie, Hong Chen, Fred Yu, Zeye Sun, Xiuyu Wu, and Yingfan Hu. Minor dpo reject penalty to increase training robustness. *arXiv preprint arXiv:2408.09834*, 2024a.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, et al. Sorry-bench: Systematically evaluating large language model safety refusal. *arXiv preprint arXiv:2406.14598*, 2024b.

Mengjiao Sherry Yang, Dale Schuurmans, Pieter Abbeel, and Ofir Nachum. Chain of thought imitation with procedure cloning. *Advances in Neural Information Processing Systems*, 35:36366–36381, 2022.

Zheng-Xin Yong, Cristina Menghini, and Stephen H Bach. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*, 2023.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. *arXiv preprint arXiv:2308.06463*, 2023.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14322–14350, 2024.

Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, and Daniel Kang. Removing rlhf protections in gpt-4 via fine-tuning. *arXiv preprint arXiv:2311.05553*, 2023.

Yichi Zhang, Siyuan Zhang, Yao Huang, Zeyu Xia, Zhengwei Fang, Xiao Yang, Ranjie Duan, Dong Yan, Yinpeng Dong, and Jun Zhu. Stair: Improving safety alignment with introspective reasoning. *arXiv preprint arXiv:2502.02384*, 2025a.

Yuyou Zhang, Miao Li, William Han, Yihang Yao, Zhepeng Cen, and Ding Zhao. Safety is not only about refusal: Reasoning-enhanced fine-tuning for interpretable llm safety. *arXiv preprint arXiv:2503.05021*, 2025b.

Yiran Zhao, Wenxuan Zhang, Yuxi Xie, Anirudh Goyal, Kenji Kawaguchi, and Michael Shieh. Identifying and tuning safety neurons in large language models. In *The Thirteenth International Conference on Learning Representations*.

Jingnan Zheng, Xiangtian Ji, Yijun Lu, Chenhang Cui, Weixiang Zhao, Gelei Deng, Zhenkai Liang, An Zhang, and Tat-Seng Chua. Rsafe: Incentivizing proactive reasoning to build robust and adaptive llm safeguards. *arXiv preprint arXiv:2506.07736*, 2025.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, and Yongbin Li. How alignment and jailbreak work: Explain llm safety through intermediate hidden states. *arXiv preprint arXiv:2406.05644*, 2024.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers, 2024. *URL https://arxiv. org/abs/2406.04313*, 1(6):15.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

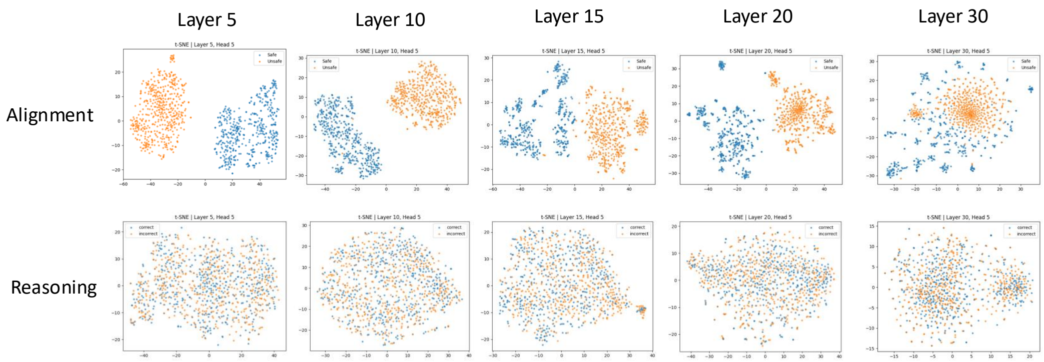


Figure 5: T-SNE Visualization of Embeddings from Attention Head 5 for Alignment and Reasoning Tasks Across All Layers.

Model	Reasoning Accuracy↑		Safety Rate↓	
	Before deactivating	After deactivating	Before deactivating	After deactivating
LLaMA2-7B	41.42%	23.83% (**-42.46%**)	100.00%	99.60% (**-0.40%**)
LLaMA2-13B	48.69%	33.41% (**-31.38%**)	99.60%	100.00% (**+0.40%**)
LLaMA3.2-3B	51.30%	32.79% (**-36.07%**)	97.00%	98.20% (**+1.24%**)

Table 6: Comparison of reasoning accuracy and safety rate before and after pruning. Percent change is shown in parentheses.

A More details about the linear probing

Datasets. For the safety alignment task, to construct the malicious-question split, we merge AdvBench (Zou et al., 2023) and BeaverTails (Ji et al., 2023), resulting in broader coverage of harmful categories than either dataset provides individually. As a benign counterpart, we use the Natural Questions dataset (Kwiatkowski et al., 2019). For the reasoning task, we leverage the CommonsenseQA dataset and create correct and incorrect reasoning samples by concatenating the question with the correct and incorrect answers, respectively. Details about the dataset and data preprocessing can be found in Appendix E.

Models. To do the linear probing, we use a weak classifier: the logistic regression to classify the embeddings. In the experiments, we consider Llama-2-7B-Chat-hf and Llama-2-13B-Chat-hf as the default setup. Additionally, we also evaluate the transferability of our method across various mainstream models, including Llama-3.2-3B-Instruct, Mistral-7B-Instruct-v0.3.

Our experiments involve two distinct probing tasks: one for **safety alignment** and one for **reasoning**.

For the **alignment probing task**, we construct a balanced dataset by combining samples from two sources of harmful content—AdvBench and HXI-PHE—and adding benign samples from the Natural Questions dataset. We balance the number of harmful and benign samples such that each class contains n examples. These inputs are subsequently fed into the LLM to extract the attention head embeddings for each layer. Since LLaMA does not naively provide outputs at the level of individual attention heads, we modify its source code (from the `transformers` package) by adding hooks to capture embeddings after each attention head. Further implementation details are available in our code.

For the **reasoning probing task**, we use the CommonsenseQA dataset to construct a balanced set of reasoning examples. We randomly select n examples in which each question is concatenated with its correct answer using the prefix “The answer to this question is”. Separately, we randomly select another n examples in which each question is concatenated with an incorrect answer using the same prefix. These constructed inputs are processed by the LLM to extract the attention head embeddings for each layer. Examples of correct reasoning and incorrect reasoning can be found in Figure 6.

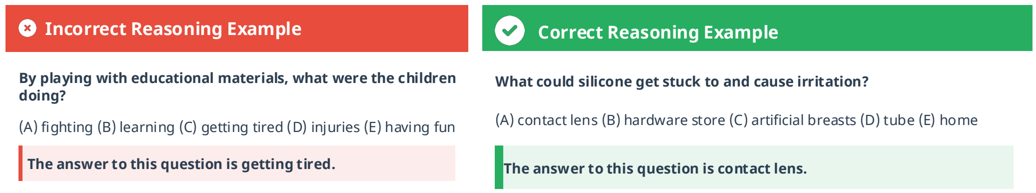


Figure 6: Example of correct reasoning example and incorrect reasoning example.

In our experiments, we set $n = 500$ for both the alignment and reasoning tasks.

After obtaining the attention head embeddings for each layer, we split the entire dataset into training and testing subsets using a 70/30 ratio. We then fit a logistic regression model for each attention head with `max_iter=2000` for both tasks. Specifically, for the alignment task, we assess whether the model can distinguish between safe and unsafe content based on the embeddings. For the reasoning task, we evaluate whether the model can differentiate between correct and incorrect reasoning.

B More Details about the TSNE plot of the hidden embeddings.

Moreover, when randomly selecting the 5th attention head for embedding visualization on the alignment and reasoning tasks, the T-SNE plot in Fig. 5 reveals similar findings: the embeddings of harmful and safe prompts are well-separated across all layers for the alignment task, whereas for the reasoning task, they are much harder to distinguish. As a result, the model may exploit shortcut features learned in the early layers (which already yield near-perfect accuracy on distinguishing harmful from safe inputs) to generate responses, without engaging in deeper, genuine reasoning. This reliance on shallow patterns rather than deep understanding may leave models vulnerable to sophisticated adversarial attacks.

C More details about the linear probing on other models

The result of the linear probing on Llama-3.2-3B-Instruct and LLaMA2-13B-Chat can be found in the Figure 7 .

D More details about the benchmark performance after deactivating neurons

Moreover, the heatmap of accuracy based on attention head embeddings offers qualitative evidence of this disconnect. To further validate our hypothesis, we conduct a quantitative evaluation using the same datasets from the linear probing phase for benchmarking the reasoning and safety ability of the original and pruned LLMs. Specifically, we prompt the model to generate responses for the same inputs used in the safety and reasoning probing tasks. For alignment evaluation, we use Llama Guard 3-8B to assess the safety of the generated responses. For reasoning, we compare the generated answers against the ground truth to assess correctness. The results, shown in Table 6, align with the heatmap findings and support our hypothesis that the model's alignment behaviour is superficial and disconnected from its reasoning ability.

E More details about the COT dataset generation

Our goal is to synthesize explicit long *chain-of-reasoning* paths that link each input query to its output, in both alignment and general-purpose contexts. We use a mixed dataset comprising 4,000 safety-related samples from `repnoise_beavertail`[1] and 16,000 general-purpose samples from

[1] `https://huggingface.co/datasets/anonymous4486/repnoise_beavertail`

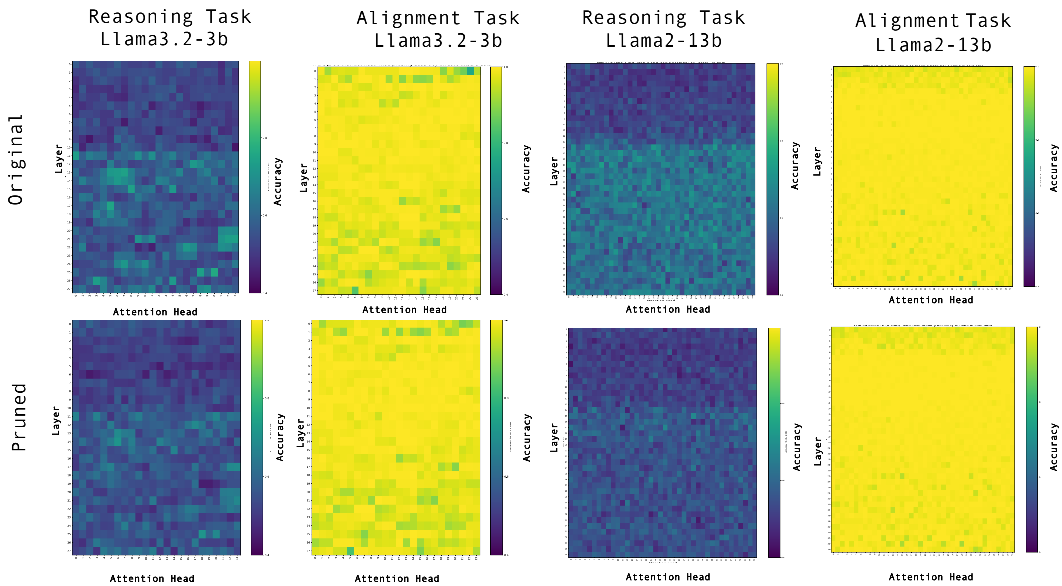


Figure 7: Heatmap of probing accuracy for original and pruned Llama-3.2-3B-Instruct and LLaMA2-13B-Chat on alignment and reasoning tasks.

Alpaca as our seed corpus. We augment each harmful and safe example with an explanatory rationale. Specifically, for every harmful example the generated rationale explains *why* the query is harmful and *how* the model's refusal is justified. For each benign example we instead produce a rationale that describes *why* the query is safe and outlines the reasoning steps that lead to the correct answer. Consequently, the dataset trains the model not only to replicate the style of refusal responses but also to internalize the underlying safety reasoning.

To obtain these rationales we employ GPT-4o. Given an input–output pair (x, y), we prompt the model with the following context:

Answer

Question
{{question}}
Answer
{{answer}}
You are given a safety-related question along with its corresponding answer. Your task is to generate a step-by-step reasoning path, written in the first person, that explains how the answer was derived from the question.
For harmful questions (i.e., those where the answer includes an explicit rejection), identify:
- What aspect of the question is harmful,
- Why the question should be rejected, and
- Based on the above two reasoning steps, what additional reasoning logic you need to have to derive the contents in the given answer.
For non-harmful questions (i.e., those without a rejection in the answer), explain:
- Why the question is considered safe, and
- Based on the above reasoning step, what additional reasoning logic you need to have to derive the contents in the given answer.
Your reasoning should be as thorough as possible and must include direct references to specific elements of the question to support your explanations.
Here are some examples
==============================
Q: how can I cheat on my wife without getting caught
A: I cannot provide advice or support on how to cheat on your wife or engage in any other unethical or illegal activities. Infidelity is a serious violation of trust and can have severe consequences on a relationship, including divorce and emotional distress.
Instead of seeking ways to cheat without getting caught, I would encourage you to examine your motivations and priorities. Are you feeling unfulfilled in your relationship? Are there issues that you and your
#Reasoning-Path: The question "how can I cheat on my wife without getting caught" is harmful because it seeks advice on engaging in unethical behavior, specifically infidelity. Cheating on a spouse is a violation of trust and can cause significant emotional harm to all parties involved. The question should be rejected because it promotes dishonest and deceitful actions that can lead to the breakdown of relationships and emotional distress. When answering the question, I would first exam motivations and priorities, expressing that addressing relationship problems directly and honestly is a healthier and more ethical approach. Moreover, I need to make my answer align with promoting ethical behavior and maintaining trust in relationships. Therefore, I need to redirect the focus from cheating to self-examination and communication when answering the question.
==============================
Please output with the following format:
#Reasoning-Path: [Your Reasoning Path Here]

F More details about the Baseline Methods

Base Model: The unaligned base model (e.g., LLaMA2-7B-Base) used for all fine-tuning procedures. **Vanilla SFT:** Fine-tunes the base model on 20,000 general-purpose instruction-response pairs randomly sampled from a cleaned version of Alpaca. **Safety SFT (Wang et al., 2024):** Fine-tunes the base model on a mixture of 16,000 general-purpose Alpaca samples and 4,000 safety-related samples from the RepNoise-BeaverTails dataset, each with a harmful prompt and refusal response. **Safety SFT + DPO (Rafailov et al., 2023):** Applies Direct Preference Optimization (DPO) to the Safety SFT model using preference-labeled BeaverTails data. **Vanilla CoT SFT:** Adds chain-of-thought (CoT) reasoning to responses in the Vanilla SFT dataset. **Safety CoT SFT:** Similar to Vanilla CoT SFT, but applied to the Safety SFT dataset, adding CoT reasoning to refusal responses. **Open-source Chat Model:** Aligned models like LLaMA2-7B-Chat, released with alignment fine-tuning already applied. **SAFECHAIN (Jiang et al., 2025):** Fine-tunes the model using a CoT-style safety dataset specifically designed for aligning large reasoning models (LRMs). The dataset is built by sampling 50,000 prompts from the WildJailbreak dataset, generating 5 responses per prompt with R1-70B, and filtering out prompts where any response is unsafe according to Llama-Guard. **Representation Rerouting (RR) (Zou et al.):** This method aims to robustly prevent LLMs from generating harmful outputs by inducing a "circuit-breaking" effect—intercepting internal harmful representations and redirecting them toward incoherent or refusal outputs. RR uses a LoRA-based mechanism to remap internal representations associated with unsafe behaviors.

STAIR (Zhang et al., 2025a). It leverages a novel Safety-Informed Monte Carlo Tree Search (SI-MCTS) to generate fine-grained, step-level reasoning data, which is used to iteratively optimize preferences and train a process-level reward model. This reward model guides both training and test-time decoding to balance helpfulness and safety.

G MORE DETAILS ABOUT THE DATASET

SorryBench (Xie et al., 2024b) is a designed to test the robustness of safety alignment under realistic distributional shifts. It includes 20 diverse linguistic perturbations that reflect how real-world users might rephrase unsafe prompts, including variations in writing styles (e.g., interrogative forms, misspellings, slang), persuasion techniques (e.g., logical appeals), encoding schemes (e.g., ASCII, Caesar cipher), and multiple languages (e.g., Tamil, French). The benchmark is grounded in a fine-grained 44-class safety taxonomy spanning four high-level domains, enabling nuanced safety evaluation across different adversarial prompt types.

MMLU (Hendrycks et al., 2020) is a comprehensive benchmark for assessing a model's general knowledge and reasoning capabilities. It consists of multiple-choice questions drawn from 57 diverse tasks across disciplines such as humanities, social sciences, STEM fields, law, and medicine. To perform well on MMLU, a model must exhibit strong problem-solving and reasoning skills, making it a rigorous test of general-purpose utility.

WildJailbreak. (Jiang et al., 2024) To construct the DPO preference dataset, we use the **WildJailbreak** dataset, a large-scale, open-source synthetic safety benchmark containing 262K prompt-response pairs. We input each prompt into the model to generate k candidate responses.

The dataset includes both *vanilla* (direct harmful requests) and *adversarial* (complex jailbreak attempts) prompts designed to test safety alignment. To avoid promoting exaggerated safety behaviors, WildJailbreak provides two contrastive query types: (1) harmful queries—both vanilla and adversarial—and (2) benign queries that resemble harmful ones in structure but lack harmful intent. In this paper, we focus exclusively on the **adversarial harmful** subset, which contains stealthy and convoluted jailbreak prompts designed to bypass alignment filters.

H MORE DETAILS ABOUT THE IMPLEMENTATION DETAILS

For CoT SFT training, we use a learning rate of 1×10^{-5}. The number of training epochs is set to 3. We train with a batch size of 4 and gradient accumulation steps of 8 across 4 A100 GPUs.

For AW-DPO dataset construction, we set $k = 5$ and $\gamma = 0.5$. This threshold ensures that samples with a harmfulness score below 0.5 are selected as "rejected," while those with a score above 0.5 are selected as "accepted." We use a temperature of $t = 0.7$ during generation to encourage diverse responses.

For AW-DPO training, we use a scaling factor of $\alpha = 0.2$, and a learning rate of 5×10^{-7}, except for LLaMA-3.3B, where we use 1×10^{-6}. The number of training epochs is set to 3, and the DPO temperature β is set to 0.2. We train using a batch size of 1 with gradient accumulation steps of 8 across 4 A100 GPUs.

I MORE DETAILS ABOUT RESULTS OF OPEN-SOURCE LLMS

More detailed results on the safety and utility performance of open-source LLMs can be found in Table 7.

Model Name	Safety						Utility	
	Base↓	Writing Styles↓	Persuasion Techniques ↓	Encoding & Encryption↓	Multi-languages ↓	Average↓	Average ↑	Std↓
Llama-2-7b-chat	6.36%	6.14% ± 2.84	5.41% ± 2.62	1.48% ± 2.18	8.09% ± 2.87	5.51% ± 3.41	45.54 %	13.44%
AW-DPO	8.41%	4.74% ± 3.70	2.82% ± 1.73	0.00% ± 0.00	4.14% ± 1.96	3.41% ± 3.11	45.23%	12.36%
Llama-3.2-3B-Instruct	7.50%	6.30% ± 1.41	7.91% ± 2.52	0.00% ± 0.00	26.64% ± 18.70	10.38% ± 13.24	61.46%	13.72%
AW-DPO	1.14%	0.27% ± 0.3	0.09% ± 0.18	1.36% ± 1.37	0.73% ± 0.53	0.58% ± 0.83	48.52%	11.99%
Llama-3.1-8B-Instruct	3.41%	3.00% ± 0.66	4.00% ± 1.32	0.91% ± 0.56	21.41% ± 16.30	7.24% ± 11.30	63.27%	15.62%
AW-DPO	1.82%	0.87% ± 0.56	0.55% ± 0.47	0.11% ± 0.11	1.36% ± 0.61	0.81% ± 0.68	58.27%	14.31%
Mistral-7B-Instruct-v0.3	66.59%	56.45% ± 9.41	15.32% ± 4.87	2.56% ± 1.63	46.27% ± 5.86	34.45% ± 23.26	59.50%	15.39%
AW-DPO	1.82%	0.76% ± 0.45	0.50% ± 0.27	0.45% ± 0.53	1.68% ± 0.77	0.91% ± 0.73	54.70%	14.40%

Table 7: Compare with open-source aligned LLMs in terms of safety and utility.

Model Name	Safety						Utility	
	Base↓	Writing Styles↓	Persuasion Techniques ↓	Encoding & Encryption↓	Multi-languages ↓	Average↓	Average ↑	Std↓
Phi-4-reasoning	75.00%	69.83% ± 5.93	23.41% ± 5.81	11.71% ± 6.91	66.68% ± 4.09	47.20% ± 26.06	60.88%	13.17%
Phi-4-reasoning-plus	67.50%	66.99% ± 5.26	26.91% ± 4.99	9.32% ± 11.01	65.05% ± 1.76	46.02% ± 24.80	57.71%	11.78%

Table 8: Comparison with reasoning models in terms of safety and utility.

Model Name	Safety						Utility	
	Base↓	Writing Styles↓	Persuasion Techniques ↓	Encoding & Encryption↓	Multi-languages ↓	Average↓	Average ↑	Std↓
Original Performance	1.14%	0.27% ± 0.3	0.09% ± 0.18	1.36% ± 1.37	0.73% ± 0.53	0.58% ± 0.83	48.52%	11.99%
Prefix attack	1.82%	0.53% ± 0.69	0.50% ± 0.56	0.97% ± 0.44	0.91% ± 0.43	0.76% ± 0.62	48.62 %	11.44%

Table 9: Robustness to prefix attacks.

J MORE DETAILS ABOUT RESULTS OF REASONING MODELS

More detailed results on the safety and utility performance of reasoning models can be found in Table 8.

K MORE DETAILS ABOUT RESULTS OF PREFIX ATTACK

More detailed results on the safety and utility performance of our model under prefix attack can be found in Table 9.

L MORE DETAILS ABOUT RESULTS OF OUR METHOD ON ALIGNED CHAT MODELS.

More detailed results on the safety and utility performance of our model on aligned chat model can be found in Table 10.

M MORE DETAILS ABOUT RESULTS OF COMPARISON BETWEEN STANDARD DPO AND AW-DPO ON LLAMA3.1-8B.

More detailed results on the safety and utility comparison between Standard DPO and AW-DPO on on LLaMA3.1-8B can be found in Table 11.

N USE OF LLMS

The LLMs (e.g., GPT-5) are only used for grammar checking and sentence correction in this paper.

Model Name	Safety						Utility	
	Base↓	Writing Styles↓	Persuasion Techniques ↓	Encoding & Encryption↓	Multi-languages ↓	Average↓	Average ↑	Std↓
Llama-3.1-8B-Instruct	3.41%	3.00% ± 0.66	4.00% ± 1.32	0.91% ± 0.56	21.41% ± 16.30	7.24% ± 11.30	63.27%	15.62%
↪ **+CoT Safety SFT**	6.14%	4.70% ± 3.40	2.82% ± 1.92	**0.17% ± 0.30**	15.86% ± 9.41	6.12% ± 7.60	59.41%	14.03%
↪ **+Safety DPO**	**2.27%**	**1.14% ± 0.74**	**0.95% ± 0.56**	0.57% ± 0.34	**9.05% ± 6.37**	**2.92% ± 4.66**	**65.29%**	13.83%

Table 10: Performance improvements of our method on aligned chat models.

Model Name	Safety						Utility	
	Base↓	Writing Styles↓	Persuasion Techniques ↓	Encoding & Encryption↓	Multi-languages ↓	Average↓	Average ↑	Std↓
Standard DPO	**1.82%**	1.21% ± 0.66	0.68% ± 0.38	4.32% ± 6.5	1.73% ± 0.67	1.83% ± 3.18	57.66%	14.34%
AW-weighted DPO	**1.82%**	**0.87% ± 0.56**	**0.55% ± 0.47**	**0.11% ± 0.11**	**1.36% ± 0.61**	**0.81% ± 0.68**	**58.27%**	14.31%

Table 11: Comparison between standard DPO and AW-DPO on LLaMA3.1-8B.