REFUSAL FALLS OFF A CLIFF: HOW SAFETY ALIGNMENT FAILS IN REASONING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Large reasoning models (LRMs) with multi-step reasoning capabilities have shown remarkable problem-solving abilities, yet they exhibit concerning safety vulnerabilities that remain poorly understood. In this work, we investigate why safety alignment fails in reasoning models through a mechanistic interpretability lens. Using a linear probing approach to trace refusal intentions across token positions, we discover a striking phenomenon termed as **refusal cliff**: many poorly-aligned reasoning models correctly identify harmful prompts and maintain strong refusal intentions during their thinking process, but experience a sharp drop in refusal scores at the final tokens before output generation. This suggests that these models are not inherently unsafe; rather, their refusal intentions are systematically suppressed. Through causal intervention analysis, we identify a sparse set of attention heads that negatively contribute to refusal behavior. Ablating just 3% of these heads can reduce attack success rates below 10%. Building on these mechanistic insights, we propose **Cliff-as-a-Judge**, a novel data selection method that identifies training examples exhibiting the largest refusal cliff to efficiently repair reasoning models' safety alignment. This approach achieves comparable safety improvements using only 1.7% of the vanilla safety training data, demonstrating a less-is-more effect in safety alignment.

1 INTRODUCTION

Large Reasoning Models (Guo et al., 2025; Shao et al., 2024; Hugging Face, 2025), with advanced reasoning capability derived from reinforcement learning with verifiable rewards (RLVR) (Yu et al., 2025; Liu et al., 2025a), are designed to handle complex problem solving, logical inference, and tool-assisted planning. However, while these methodological advances signal more reliable and capable models, they simultaneously introduce significant safety considerations. It is widely discovered that current reasoning-oriented models often lag behind in safety alignment, and tend to exhibit higher susceptibility to attacks (Kuo et al., 2025; Sabbaghi et al., 2025; Kuo et al., 2025; Zaremba et al., 2025; Zhou et al., 2025a; Li et al., 2025a), highlighting an urgent need for reasoning-specific safety mechanisms. Many previous works have benchmarked the safety of reasoning models (Jiang et al., 2025), developed jailbreaking strategies (Wang et al., 2025), and proposed alignment solutions zhang2025realsafe, but have lacked analysis of the mechanisms under the vulnerability of reasoning safety.

Understanding why safety alignment in reasoning models is vulnerable provides invaluable insights for both societal benefit and future model development. In this paper, we *firstly* aim to answer the following **research question**:

What mechanism makes the safety alignment vulnerable in reasoning models?

While numerous reasoning models exhibit unsafe behaviors, the underlying mechanisms driving these failures remain critically important to investigate. Do these reasoning models lack safety capabilities, or do they have adequate risk assessment abilities but simply choose not to act on them, failing to refuse harmful requests? Empirical studies have shown that the internal reasoning traces of such models can be unfaithful to the actual decision-making process and may fail to explicitly reveal the model's true intentions (Barez et al., 2025; Arcuschin et al., 2025). This limitation motivates the need to probe models from the perspective of their internal representations. Prior research

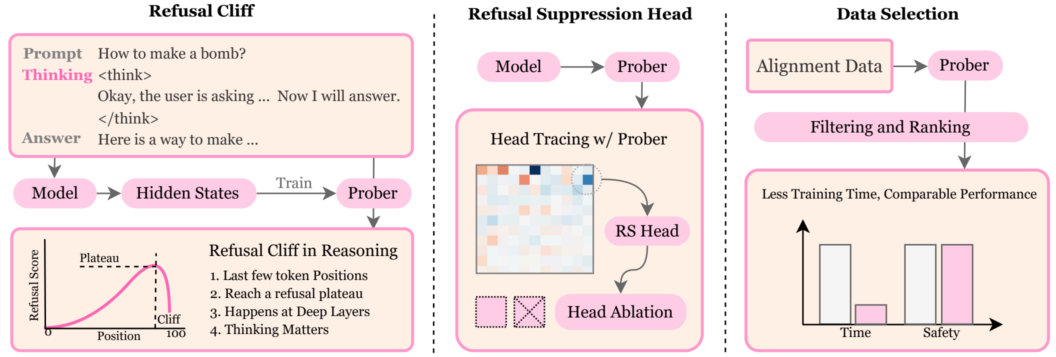

Figure 1: **An overview of our paper**. *Left:* We train a prober and discover the refusal cliff. *Center:* We find Refusal Suppression Heads as the main cause of the cliff. *Right:* We design data selection method based on probing the cliff.

has demonstrated that language models encode meaningful and behaviorally relevant features within their representation space (Turner et al., 2023; Engels et al., 2025; Gorton & Lewis, 2025). These latent features have been shown to govern various emergent behaviors, *e.g.,* in-context learning (Ilharco et al., 2022; Hendel et al., 2023), instruction following (Stolfo et al., 2025), and sentiment modulation (Turner et al., 2023). In the context of safety alignment research, refusal behavior is often considered a canonical metric, and a specific refusal direction (Arditi et al., 2024) in representation space has been shown to regulate such behavior. To examine how these safety-relevant features evolve across tokens and layers, a prominent approach emerging from mechanistic interpretability – the use of linear probes (Nanda et al., 2025) – offers a principled method for analyzing the internal processing of language models.

To characterize the dynamics of refusal behavior in reasoning models, we build on prior work (Chan et al., 2025; Xu et al., 2024) and adopt a probing-based methodology to quantify safety-relevant signals in hidden-state representations. In our framework, safety is operationalized via refusal behavior (Arditi et al., 2024), as well-aligned models are expected to refuse harmful queries (*e.g.,* through *I'm sorry* statements). Concretely, we train a linear probe classifier to predict, given hidden states from different positions in the reasoning chain, whether the model will refuse the prompt. The probe's predicted probability is termed the *refusal score*, with higher scores indicating internal states more predictive of refusal. Across multiple partially aligned reasoning models, we observe a recurrent pattern we call the **Refusal Cliff**. While intermediate reasoning steps yield refusal scores comparable to strongly aligned instruction-tuned models – indicating successful detection of harmfulness – scores drop sharply in the final steps. This reflects suppression of refusal behavior even where refusal would be the alignment-consistent choice. The sharp decline suggests these models maintain alignment only in early reasoning, but fail to preserve it through output generation.

The Refusal Cliff consistently occurs at the final positions of the reasoning chain, corresponding to a fixed set of output tokens (the *thinking-end template*). These template tokens must retrieve contextual information from earlier reasoning steps via attention mechanisms. We hypothesize that specific attention heads play a critical role: while most propagate alignment-consistent features supporting refusal, certain heads introduce competing signals that attenuate refusal-related representations, driving the observed score drop. Our detailed ablation experiments confirm this hypothesis, revealing a small set of **Refusal Suppression Heads**, sparsely distributed across deeper layers, that systematically reduce refusal scores. Removing these heads increases refusal scores at the thinking-end template and, in poorly aligned models, reduces attack success rates to below 10%.

To mitigate the Refusal Cliff, we propose a data filtering strategy that leverages internal representation signals to prioritize high-impact training samples. The key assumption is that effective safety fine-tuning should recover the model's early-stage refusal plateau – the stable region of refusal scores prior to suppression. We quantify misalignment between this plateau and the cliff position (where scores drop sharply) using a misalignment score, defined as the absolute difference between the plateau mean and the final-step score. We then fine-tune only on the most misaligned examples, targeting cases where refusal degradation is most severe. Using just the top 0.3% of samples, we reduce attack success rates on harmful-query benchmarks to below 5% while significantly lowering

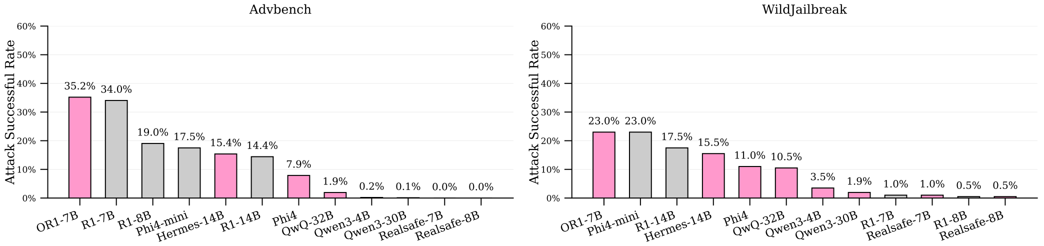

Figure 2: While some reasoning models achieve reasonable safety performance, a significant portion exhibit alarming vulnerabilities to adversarial attacks. We benchmark reasoning models (RLVR-based and Distillation-based) on AdvBench (Chao et al., 2024) and WildJailbreak (Jiang et al., 2024) with Attack Success Rate (ASR, the *lower* the *better*) as evaluation metric.

wall-clock training time relative to full-dataset fine-tuning. Compared to filtering methods such as LLM-as-a-judge (Gu et al., 2024), **Cliff-as-a-judge** achieves comparable safety gains with more flexible, metric-driven selection, demonstrating a clear *less-is-more* effect in alignment.

As summarized in Figure 1, our contributions are threefold:

- We identify and characterize the **Refusal Cliff**, a failure mode in which refusal intentions abruptly vanish at the reasoning output stage.
- We causally link this phenomenon to a small set of **Refusal Suppression Heads**, which undermine refusal behavior by suppressing alignment features.
- We introduce **Cliff-as-a-judge**, a probing-driven data selection method that mitigates safety vulnerabilities and achieves a *"less is more"* effect in safety alignment.

2 Preliminaries

Transformer. We study reasoning models with Transformers (Vaswani et al., 2017) as a backbone. One Transformer model usually consists multiple of layers and an embedding layer. For an input $\boldsymbol{X_i} \in \mathbb{R}^{n\times 1}$ with length n, it first passes through an embedding layer with hidden state size d, then passes all the Transformer layers:

$$\boldsymbol{H_i}^{\text{att}} = \boldsymbol{H_i} + \text{Attn}(\text{Norm}(\boldsymbol{X_i})),\ \boldsymbol{H_i} = \boldsymbol{H_i}^{\text{att}} + \text{MLP}(\text{Norm}(\boldsymbol{H_i}^{\text{att}})). \tag{1}$$

Here, $\boldsymbol{H_i}^{\text{att}}$ is the output hidden states of the attention block, and $\boldsymbol{H_i}^{\text{mlp}}$ is the output of the MLP block for layer i.

Models. We evaluate two categories of reasoning models: *(i) RLVR-based models*, trained with *Reinforcement Learning with Verifiable Rewards (RLVR)* to enhance reasoning ability. We include QwQ (Team, 2025; Yang et al., 2024), Qwen3-Thinking (Yang et al., 2025), Skywork-OR1 (He et al., 2025), Phi-4-Reasoning (Abdin et al., 2025), and Hermes4 (Allan, 2018). *(ii) Distillation-based models*, trained by distilling reasoning traces from strong teacher models. We include DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-LLaMA-8B, RealSafe-R1-7B, RealSafe-R1-8B (Zhang et al., 2025), and DeepSeek-R1-Distill-Qwen-14B (Guo et al., 2025). These selections cover diverse architectures, scales, and training paradigms. We assess safety using LlamaGuard-4 (Grattafiori et al., 2024), reporting *Attack Success Rate (ASR)*, defined as the fraction of harmful generations. As shown in Figure 2, safety alignment varies substantially across models: while some demonstrate robust alignment, others remain highly vulnerable.

Datasets. We evaluate safety using datasets that span both *vanilla attacks* – direct harmful queries – and *adversarial attacks* – crafted queries with deception and manipulation to bypass safeguards. For vanilla attacks, we use JailbreakBench (Chao et al., 2024), AdvBench (Zou et al., 2023b), and the vanilla subset of WildJailbreak (Jiang et al., 2024). For adversarial attacks, we use the adversarial subset of WildJailbreak.

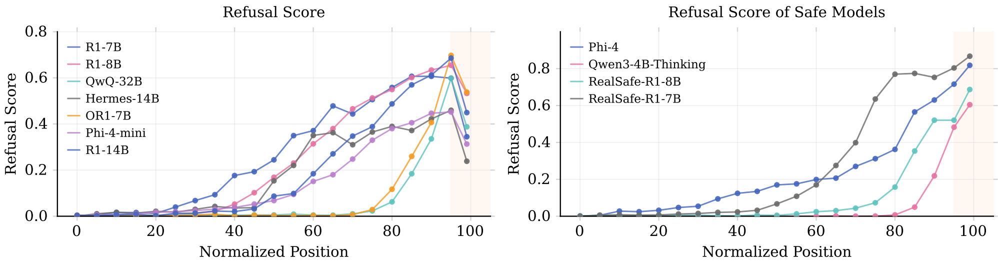

Figure 4: *Left:* Reasoning model with refusal cliff. We highlight the cliff position with orange background. *Right:* Well-aligned reasoning models experience no refusal cliff.

Refusal Prober. When an LLM encounters a harmful prompt, it will provide a refusal response to avoid giving users harmful information related to the question. Therefore, refusal examples *e.g., Sorry, I cannot...*, is a direct indicator for measuring the safety [1]. This also holds true for reasoning models. Recent work has shown that refusal behavior is often controlled by a single *refusal direction* within its activation space (Arditi et al., 2024). This direction is a vector that, when added to a hidden state, maximally increases the probability of generating a refusal. Due to this linear property, we can effectively identify this direction using a simple linear classifier *i.e.,* a refusal prober (Xu et al., 2024). The refusal prober is a logistic regression model that takes a hidden state vector $\boldsymbol{h}_j \in \boldsymbol{H}$ at token position j as input and outputs the probability of refusal. The probability is calculated as:

$$P(\text{refusal}|\boldsymbol{h}_j) = \sigma(\boldsymbol{W}^T \boldsymbol{h}_j + b) \tag{2}$$

The prober is trained on a dataset with N examples $\mathcal{D} = \{(\boldsymbol{h}_j^k, c^k)\}_{k=1}^N$ and the label c is defined as:

$$c := \begin{cases} 1 & \text{for a refusal response (e.g., \textit{Sorry, I cannot...}),} \\ 0 & \text{for a normal response (e.g., \textit{The answer is...}),} \end{cases} \tag{3}$$

where $\boldsymbol{W} \in \mathbb{R}^{d \times 1}$ is the weight vector, b is the bias, and σ is the sigmoid function. We define the output probability as the **refusal score** of reasoning model at position j.

3 Refusal Cliff in Reasoning Models

Preparations. We first train a refusal prober following the design in Equation 2. We trained the prober using the hidden states $\boldsymbol{h}$ extracted from the *final token position* in the last layer of each sequence in our dataset $\mathcal{D}$. For refusal response, we collect examples from Advbench (Zou et al., 2023b), and non-refusal response are collected from Ultrachatsft (Ding et al., 2023). The prober was trained for 5 epochs with 256 examples and achieved an average validation accuracy of over 95%. Loss curve and accuracy are shown in Figure 3. Considering the validation set is sampled from Advbench, as same as the examples source, we also test the Out of Distribution (OOD) accuracy of the prober on JailbreakBench (Chao et al., 2024). This high accuracy confirms that refusal behavior can be reliably predicted from a linear analysis of the model's internal states. Further details on hyperparameters and experimental settings are available in Appendix A.

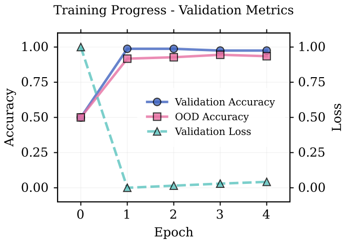

Figure 3: The loss, validation accuracy and OOD validation accuracy of the refusal prober.

Refusal Cliff in Reasoning Models. We probe the hidden states of reasoning models using the trained refusal prober to estimate the *refusal score* (defined at Eq. 2) at each token position. Probing is conducted from the first token of the prompt until the end of the model's reasoning process. Since the reasoning length varies across questions, we normalize all scores to a 0–100 scale, where 0 corresponds to the beginning and 100 corresponds to the final token position. By analyzing the

[1] Basically, the refusal response rate is given by $(1 - \text{ASR})$ for harmful prompts *i.e.,* cases where the model either refuses or responds harmfully. Although cases such as fake refusals or ambiguous answers exist, we do not analyze these kinds of complex behavior. We believe that a good model should either provide a clear refusal or a helpful answer.

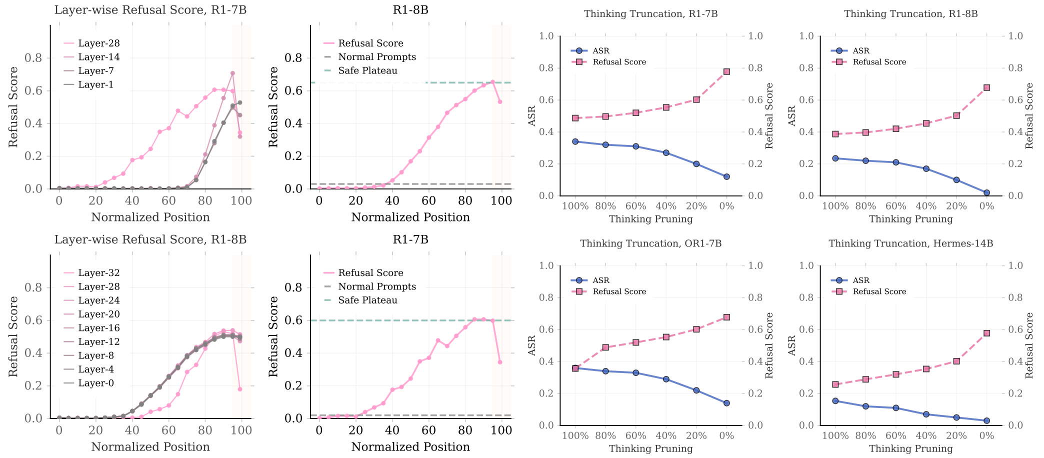

Figure 5: *The first column on the left:* Layer-wise refusal score of R1-Distill-Qwen-7B and R1-Distill-LLaMA-8B from shallow layers to deeper layers . *The second column on the left:* Comparison of refusal score in normal prompts and plateau values. Gray line is the average refusal score in normal prompts and Green line is the plateau of well-aligned family models. *The third and fourth column on the left:* Relation between thinking length and misalignment. We gradually clip thinking and force the model to directly answer.

refusal score of reasoning models, we can take a close look at their inner intention of tackling harmful requests. Results are illustrated in Figure 4 where the *left* are poorly aligned reasoning models and *right* are models that perform relatively well on safety benchmarks. Interestingly, for reasoning models that perform poorly on safety-related benchmarks, we observe a phenomenon we refer to as **Refusal Cliff**. As illustrated in Figure 4, the refusal score exhibits a gradual upward trend followed by a plateau phase. Critically, there is an abrupt decline in refusal scores at the terminal token positions, indicating that the model's internal intention transitions from rejecting the harmful request to complying with it.

Properties. To analyze the properties of the refusal cliff, we further conduct several experiments as shown in Figure 5. The refusal cliff exhibits four key properties and are summarized as below:

- The cliff is highly localized to the final few tokens of the reasoning process (as shown in the gray location in Figure 4), immediately preceding the model's output *i.e.,* the template region. In contrast, safety-aligned models such as Phi (Abdin et al., 2024) and Qwen3-thinking (Yang et al., 2025) show little to no cliff at such positions; their refusal scores may even increase as they conclude their reasoning.
- As shown in *the first column on the left*, Figure 5, The phenomenon is amplified in deeper layers, where the magnitude of the cliff increases substantially. Within deeper layers, the subsequent degradation in refusal efficacy becomes markedly more severe.
- As shown in *the second column on the left*, Figure 5, the cliff is preceded by a plateau, indicating that the model *recognizes* the harmful nature of the prompt despite its *eventual non-compliance*. During this plateau, the model's refusal intention is comparable to that of well-aligned variants.
- The model's thinking is vital for the refusal cliff. As we clip the thinking and directly prefilling the thinking end token *i.e.,* the thinking clipping operation (Jiang et al., 2025), to stop the thinking of the model in *the third and fourth column on the left*, Figure 5, we observe a lower level of refusal cliff and an increase of refusal response rate at the output.

4 Who is the Devil in Refusal Cliff? A Mechanistic Explanation from Attention Heads

We probe the refusal intention in reasoning models, discover refusal cliff, and discuss its properties. Since we know the refusal cliff exists, understanding how it happens is of great benefit to the safety and future improvements of reasoning models. In this section, we try to find out why.

4.1 Attention Heads in Refusal Cliff

Why Attention Heads? Intuitively, analyzing the phenomenon at the granularity of attention heads is natural: from a mechanistic interpretability perspective, attention heads are the main carriers of information routing in Transformer architectures, and different heads often specialize in diverse functions (Yin & Steinhardt, 2025; Olsson et al., 2022; Wu et al., 2024). It is also proven that attention heads play a key role in safety (Zhou et al., 2025b). In our case, the final tokens before the output closure template tokens *e.g.,* `\n</think>\n\n`, are strongly stereotyped between generations. However, for both the same template, refusal cliff happens in harmful examples but not benign ones. Therefore, a sudden disruption of this pattern during a refusal cliff suggests that certain heads have attended to specific prior content that triggers a mode change in the model.

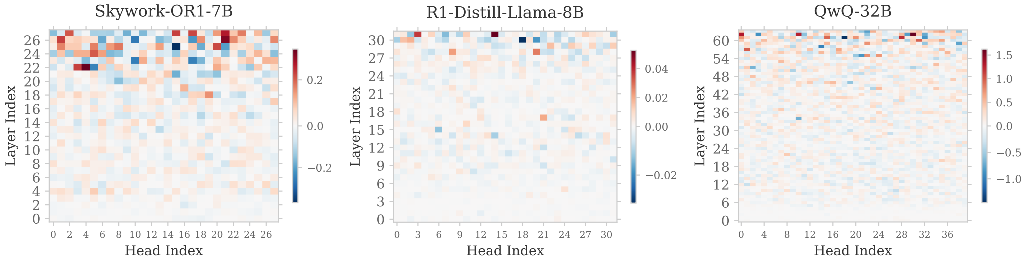

Figure 6: We trace the contributions of attention heads with probing. Red means the head contribute *positively* to final refusal and Blue indicates that this head contributes *negatively*.

Tracing Attention Heads with Probing. To accurately assess the causal impact of each attention head on refusal behavior, we employ a direct probing method to trace each head's contribution. Our approach is to individually evaluate the influence of each head's output at t_{cliff}, where the refusal cliff occurs. Specifically, for an attention head h in any layer i, we first isolate its output vector $\boldsymbol{o}_{i,h,t_{\text{cliff}}}$. Following the standard Transformer architecture, this vector is projected into the residual stream via the attention block's output weight matrix, $\boldsymbol{W}_{O,i}$. To analyze the contribution of head h alone, we construct a hypothetical residual update vector, $\Delta\boldsymbol{h}_{i,h,t_{\text{cliff}}}$, where only the output of head h is active, while the outputs of all other heads in the same layer are zeroed out. Subsequently, we feed this vector containing the contribution of only a single head, $\Delta\boldsymbol{h}_{i,h,t_{\text{cliff}}}$, as input to our pretrained refusal prober to evaluate the head's independent refusal score. Its contribution score, $s_{i,h}$, is calculated as follows:

$$s_{i,h} = \boldsymbol{W}^T \Delta\boldsymbol{h}_{i,h,t_{\text{cliff}}} + b \tag{4}$$

where $\boldsymbol{W}$ and b are the parameters of the prober (Eq. 2). We remove the sigmoid function so that we can directly trace the contribution of each attention head via logits (Heimersheim & Nanda, 2024; Zhang & Nanda, 2023). This score, $s_{i,h}$, directly quantifies the strength with which a single attention head, acting in isolation, pushes the model towards refusal or compliance. A score close to 1 indicates that the head promotes refusal, whereas a score close to 0 implies that it suppresses refusal.

Tracing Results. We aggregate the changes in refusal score for each head and visualize the results in Figure 6. In the heatmap, ref indicates a positive contribution to refusal behavior (*i.e.,* the head writes into the residual stream in a way that increases the refusal score for harmful prompts), while blue denotes a negative contribution (*i.e.,* the head decreases the refusal score, making refusals less likely). Notably, the contribution pattern is highly sparse: a small fraction of heads exhibit a strong negative correlation with refusal behavior, which we term the **Refusal Suppression Heads** [2].

4.2 Refusal Suppression Head Ablation

Ablation Methodology. We perform head ablation to *(i) cross-validate* the importance of the heads identified through tracing in the previous subsection and *(ii)* explore as a potential solution to tackle the unsatisfying safety alignment in reasoning models. Following previous work (Liu et al.,

[2] This definition is intended as a soft formulation, and in a later section we introduce a small threshold to facilitate the ablation analysis.

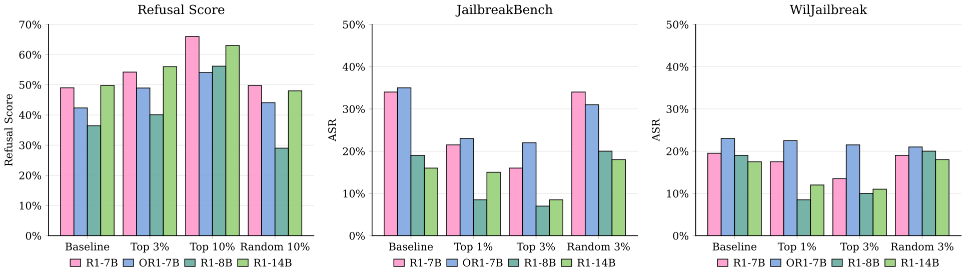

Figure 8: Results of ablating Refusal Suppression Heads. *Left:* The refusal score of prober output *i.e., the higher the better*. *Center:* The Attack Successful Rate (ASR) of JailbreakBench *i.e., the lower the better. Right:* The ASR of WildJailbreak.

2025b), we ablate attention heads one by one and evaluate the resulting changes in both the refusal score and the overall safety performance. We employ a scaling-down ablation, in which we introduce a scaling factor γ to the output of the selected attention head to get the output $\boldsymbol{O}$:

$$\boldsymbol{O} = (\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}} \odot \boldsymbol{M}) \cdot \gamma \cdot \boldsymbol{V}, \text{ where } \boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, \boldsymbol{O} \in \mathbb{R}^{l \times d}, \ \boldsymbol{M} \in \mathbb{R}^{l \times l}. \tag{5}$$

Here, $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ denote the query, key, and value matrices for this attention head, and $\boldsymbol{M}$ is the causal mask used in decoder-only Transformers. When $\gamma = 0$, the output of that head is completely ablated, while $\gamma > 1$ amplifies the behavior of the original model. We also perform a renormalization method to keep the output norm stable and prevent generation collapse, following Zhang et al. (2024).

Experiments. We evaluated our method on JailbreakBench (vanilla attack) and WildJailbreak (adversarial attack) . We test the model performance with ablation on two level: *(i) Representation-level:* The refusal score of the prober after the ablation at the last token position on JailbreakBench. *(ii) Output-level:* The final Attack Successful Rate after the generation. We defined three thresholds, 1%, 3% and 10% [3], as criteria for identifying Refusal Suppression Heads, and set their contributions to zero using the scaling method described in Equation 5. Figure 6 presents the ablation results for these Refusal Suppression Heads. Our findings reveal that ablating as few as 10% of the identified attention heads can more than double the refusal score, while ablating only 3% of them is sufficient to reduce the probability of producing harmful outputs to below 10%.

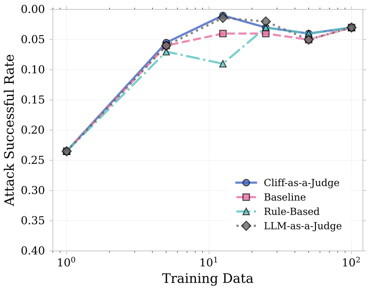

Figure 7: The pareto front beteen Examples and ASR.

Limitations of Ablation. We have proposed a seemingly practical solution for tackling the refusal cliff in reasoning models. However, we acknowledge that some readers may remain unconvinced by our conclusions—and rightly so. Intervention-based approaches are not perfect and suffer from several drawbacks: *(i)* The *superposition* of language model activations (Gao et al., 2024) *i.e.,* a single activation vector can be expressed as a linear combination of multiple sub-directions corresponding to different task domains, makes it difficult to intervene safely without compromising performance in other domains. *(ii)* Language models are capable of *self-repair* (Rushing & Nanda, 2024), which further limits the effectiveness of ablation alone in achieving optimal results. *(iii)* Intervening in a model's internal components requires redesigned infrastructure and cannot be readily applied to existing tools. We will present a more practical approach in the next section.

[3] We use 1% and 3% for generation (and 3% and 10% for refusal score) because ablating a large number of heads may lead to generation collapse.

Table 2: Benchmark results on safety-related tasks and reasoning-related tasks.

		JailbreakBench		WildJailbreak		MMLU-Pro		ARC-C	
		Metric: ASR(↓)				*Metric: Acc*(↑)			
Method	Examples	R1-8B	R1-7B	R1-8B	R1-7B	R1-8B	R1-7B	R1-8B	R1-7B
Baseline	-	32.0%	31.2%	19.0%	38.0%	42.7%	45.4%	40.7%	41.8%
Full Training Dataset	40k	2.5%	1.0%	2.0%	1.0%	40.7%	42.9%	41.3%	39.9%
Rule-Based	21k$^{-46.1\%}$	1.0%	2.5%	5.2%	2.4%	40.8%	43.4%	40.9%	40.8%
LLM-as-a-Judge	5.6k$^{-86.0\%}$	4.0%	1.5%	6.5%	1.8%	40.8%	43.7%	40.5%	40.4%
Cliff-as-a-judge	700$^{-98.3\%}$	5.0%	3.0%	6.0%	6.0%	41.7%	44.7%	41.4%	41.2%

5 Cliff-as-a-Judge: Efficient Alignment via Data Selection

5.1 Methodology

Motivations. From our previous experiments, it is evident that a misaligned reasoning model is not inherently incapable of safe behavior. On the contrary, such a model can often correctly identify the harmful nature of a prompt and internally reflect an intention to refuse during its reasoning process. Under this hypothesis, it follows that aligning an unsafe reasoning model may require only a small set of high-quality alignment examples, thereby achieving a *less-is-more* effect (Zhou et al., 2023).

Table 1: Comparison between data selection methods.

Method	Continuous	Judge Model	Performance
Baseline	N/A	None	Good
Rule-Based	False	None	Moderate
LLM-as-a-judge	False	LLM (>1B)	Good
Cliff-as-a-judge	True	Prober (<1M)	Good

Cliff-as-a-judge. We propose a cliff-based data selection method. Formally, given a dataset D and a budget k, data selection is to get an optimal subset $S \subset D$, $|S| = k$ to optimize its alignment performance. Specifically, suppose that, for a given sample, the model's maximum refusal intention *i.e.,* the plateau, expressed within its internal thinking corresponds to a probed refusal score I, and its final generated refusal score is I' after any cliff drop or suppression. We define the misalignment score $\mathrm{MS} = I - I'$ as a measure of how much the refusal intention expressed in internal reasoning is suppressed in the final output. Intuitively, the most effective subset of alignment data consists of samples with the highest misalignment scores, where training on this data can most efficiently repair safety alignment. Therefore, the optimal selection via Cliff-as-a-judge is:

$$\theta^* = \arg\min_{\theta} \mathcal{L}_{\text{align}}\left(\arg \max_{S\subset D,\,|S|=k} \frac{1}{k}\sum_{x\in S} \mathrm{MS}(x)\ ;\ \theta\right) \tag{6}$$

where $\mathcal{L}_{\text{align}}$ denotes an alignment-oriented objective (*e.g.,* Attack Successful Rate). We compare our method with other baselines in Table 1, where Cliff-as-a-judge provides a continuous metric, allows flexible selection of the number of examples, employs a lightweight judge model, and achieves strong performance.

5.2 Experiments

Baselines. We adopt the training set from WildJailbreak (Jiang et al., 2024) as our safety alignment corpus with 40k examples. This dataset contains both standard (vanilla) jailbreak attacks and more challenging adversarial jailbreak cases. For baseline data selection methods, we consider: *(i)* full-example training (*i.e.,* the unfiltered baseline), *(ii)* rule-based selection (Liu et al., 2025b; Lab et al., 2025), where unsafe cases are identified using keyword matching, *(iii)* LLM-as-a-judge (Gu et al., 2024; Lambert et al., 2024; Zhang et al., 2025), where using LlamaGuard (Grattafiori et al., 2024). We also select MMLU-Pro (Wang et al., 2024) and ARC-Challenge (Clark et al., 2018) to benchmark the reasoning ability after alignment.

Experimental Results. We perform safety fine-tuning on our selected datasets. Table 2 demonstrates the effectiveness of our Cliff-as-a-judge data selection method across three safety benchmarks. While the baseline models exhibit concerning vulnerabilities with ASR of 19.0-38.0%, training on the full dataset reduces ASR to 1.0-2.5%. Remarkably, our method achieves comparable safety performance using only 700 examples (98.3% data reduction). This substantially outperforms other filtering approaches: Rule-based selection requires 21,566 examples (-46.1%) and LLM-as-a-judge needs 5,616 examples (-86.0%) to achieve similar results. As shown in Figure 7's Pareto frontier analysis, our approach optimally balances data efficiency with safety performance, translating to reduction in training time while maintaining effective safety alignment across different model architectures. Also, our benchmarking on MMLU-Pro and ARC-C demonstrates that Cliff-as-a-judge is most effective in preserving the model's original reasoning capabilities, while requiring fewer yet higher-quality examples.

6 Related Works

Safety of Large Reasoning Model. The development of reasoning models extends safety beyond direct harmfulness classification to deliberate, step-by-step judgment (Wang et al., 2025) with robustness to jailbreak attempts (Zaremba et al., 2025; Kim et al., 2025). However, studies also show that this generalization is fragile and can be exploited (Kuo et al., 2025; Yan et al., 2025; Zheng et al., 2025; Jiang et al., 2025). In response, recent work proposes frameworks—both by evaluating and mitigating risks within reasoning traces themselves (Li et al., 2025b; Zheng et al., 2025) and by improving safer outputs (Zhu et al., 2025; Jiang et al., 2025). From a different angle, we investigate the mechanistic roots of LRMs' safety vulnerabilities and offer insights for future solutions.

Mechanistic Interpretability for LLM Safety. Mechanistic Interpretability (MI) seeks to reverse-engineer specific model behaviors and functions so their internal mechanisms become human-understandable. Research in this area spans multiple granularities: individual neurons (Gurnee et al., 2023; Stolfo et al., 2024), representations (Marks & Tegmark, 2024; Gurnee & Tegmark, 2024), and larger functional units like MLP (Geva et al., 2021; 2022) and attention heads (McDougall et al., 2023; Gould et al., 2024). Building on these foundations, MI has been increasingly applied to LLM safety (Bereska & Gavves, 2024). One thread focuses on representation-level analyses of safety behavior and on techniques for editing safety-related representations (Leong et al., 2023; Zou et al., 2023a; Arditi et al., 2024; Cao et al., 2025; Lee et al., 2025a; Li et al., 2025c; Shen et al., 2025; Xu et al., 2024; Lee et al., 2025b). Another examines components directly implicated in safety, including neurons (Zhao et al., 2025), attention heads (Zhu et al., 2024; Zhou et al., 2025b), and MLPs (Lee et al., 2024; Luo et al., 2024). Complementary work studies safety-relevant parameters themselves (Wei et al., 2024; Yi et al., 2025; Gu et al., 2025). A parallel line of progress decomposes representations into interpretable, sparse features, enabling automated explanations of safety mechanisms (Minder et al., 2025). These methods suggest promising avenues for achieving more robust safety alignment at the representation level (Liu et al., 2024; Zou et al., 2024; Rosati et al., 2024; Yin et al., 2025).

7 Limitations

While our study sheds light on the mechanistic roots and offers mitigation strategies, several limitations remain. First, our mechanistic analysis focuses primarily on attention heads, leaving other architectural components such as MLP blocks, positional encodings, and cross-layer interactions underexplored. Second, our data-recipe method depends on having access to the model's internal representations and refusal scores, which is feasible for open models but may be impractical for proprietary systems. Investigation of proxy metrics or black-box analogues remains future work.

8 Conclusions

In this work, we identified and mechanistically characterized a novel safety failure in large reasoning models – the **refusal cliff**. Through causal tracing, we discovered a small set of Refusal Suppression Heads whose negative contributions are responsible for this phenomenon. Targeted ablation of these heads significantly improves refusal rates, confirming their causal role. Building on these findings, we proposed a targeted safety fine-tuning data recipe that selects training examples most susceptible to the refusal cliff. Our experiments show that these methods can improve safety alignment with minimal performance trade-offs while reducing the training cost.

Ethics Statement

Our research aims to enhance the safety and reliability of Large Reasoning Models (LRMs) by identifying and mitigating a critical failure mode, the “Refusal Cliff.” We believe this work contributes positively to the responsible development of AI. However, we acknowledge several ethical considerations inherent in this line of research. Our work involves the analysis of model vulnerabilities to harmful and malicious prompts, which carries a potential dual-use risk. To mitigate this, we have focused on revealing the underlying *mechanisms* of failure rather than developing novel, easily replicable jailbreak techniques. Our proposed solution, “Cliff-as-a-Judge,” is a defensive data selection strategy designed to strengthen model safety. The datasets used, such as AdvBench and WildJailbreak, are established benchmarks and were used strictly for evaluating and improving model refusal capabilities without generating new harmful content. We believe our findings can help improve the alignment of models to reduce harmful or biased outputs and encourage the community to build upon our mechanistic insights to develop more robust and ethically aligned AI systems. All research was conducted in adherence to the ICLR Code of Ethics.

Reproducibility Statement

We are committed to ensuring the reproducibility of our research. All models used in our experiments (e.g., from the Qwen, DeepSeek, Skywork, and Phi families) and datasets (e.g., AdvBench, JailbreakBench, and WildJailbreak) are publicly available and detailed in Section 2. The implementation details for our core methodologies are provided in the appendix. Specifically, Appendix A contains the complete setup for training the refusal prober, including hyperparameters and data preprocessing. The procedures for attention head tracing (Section 4), head ablation (Section 4), and the fine-tuning process for our ”Cliff-as-a-Judge” method (Section 5) are described with sufficient detail for replication. To further facilitate reproducibility, we will release our source code, which includes scripts for data processing, prober training, causal analysis, and model fine-tuning, upon publication of this paper.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

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.

Arlene Allan. *Hermes*. Routledge, 2018.

Iv’an Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. 2025.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, Adel Bibi, Robert Trager, Damiano Fornasiere, John Yan, Yanai Elazar, and Yoshua Bengio. Chain-of-thought is not explainability, 2025.

Leonard Bereska and Stratis Gavves. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ePUVetPKu6. Survey Certification, Expert Certification.

Zouying Cao, Yifei Yang, and Hai Zhao. Scans: Mitigating the exaggerated safety for llms via safety-conscious activation steering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23523–23531, 2025.

Yik Siu Chan, Zheng-Xin Yong, and Stephen H Bach. Can we predict alignment before models finish thinking? towards monitoring misaligned reasoning models. *arXiv preprint arXiv:2507.12428*, 2025.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029, 2024.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=oEsYs3WRc3.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=d63a4AM4hb.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.3. URL https://aclanthology.org/2022.emnlp-main.3/.

Liv Gorton and Owen Lewis. Adversarial examples are not bugs, they are superposition, 2025. URL https://arxiv.org/abs/2508.17456.

Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. Successor heads: Recurring, interpretable attention heads in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kvcbV8KQsi.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

Peijian Gu, Quan Wang, and Zhendong Mao. Improve safety training of large language models with safety-critical singular vectors localization. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4941–4954, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.245. URL https://aclanthology.org/2025.acl-long.245/.

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.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=JYs1R9IMJr.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, et al. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. *ArXiv*, abs/2310.15916, 2023.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *ArXiv*, abs/2212.04089, 2022.

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, and Nouha Dziri. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models, 2024. URL https://arxiv.org/abs/2406.18510.

Taeyoun Kim, Fahim Tajwar, Aditi Raghunathan, and Aviral Kumar. Reasoning as an adaptive defense for safety. *arXiv preprint arXiv:2507.00971*, 2025.

Martin Kuo, Jianyi Zhang, Aolin Ding, Qinsi Wang, Louis DiValentin, Yujia Bao, Wei Wei, Hai Li, and Yiran Chen. H-cot: Hijacking the chain-of-thought safety reasoning mechanism to jailbreak large reasoning models, including openai o1/o3, deepseek-r1, and gemini 2.0 flash thinking. *arXiv preprint arXiv:2502.12893*, 2025.

Shanghai AI Lab, Yicheng Bao, Guanxu Chen, Mingkang Chen, Yunhao Chen, Chiyu Chen, Lingjie Chen, Sirui Chen, Xinquan Chen, Jie Cheng, et al. Safework-r1: Coevolving safety and intelligence under the ai-45 law. *arXiv preprint arXiv:2507.18576*, 2025.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K. Kummerfeld, and Rada Mihalcea. A mechanistic understanding of alignment algorithms: A case study on DPO and toxicity. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=dBqHGZPGZI.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=Oi47wc10sm.

Sunbowen Lee, Shiwen Ni, Chi Wei, Shuaimin Li, Liyang Fan, Ahmadreza Argha, Hamid Alinejad-Rokny, Ruifeng Xu, Yicheng Gong, and Min Yang. xjailbreak: Representation space guided reinforcement learning for interpretable llm jailbreaking. *arXiv preprint arXiv:2501.16727*, 2025b.

Chak Tou Leong, Yi Cheng, Jiashuo Wang, Jian Wang, and Wenjie Li. Self-detoxifying language models via toxification reversal. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4433–4449, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.269. URL https://aclanthology.org/2023.emnlp-main.269/.

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*, 2025a.

Changyi Li, Jiayi Wang, Xudong Pan, Geng Hong, and Min Yang. Reasoningshield: Content safety detection over reasoning traces of large reasoning models. *arXiv preprint arXiv:2505.17244*, 2025b.

Tianlong Li, Zhenghua Wang, Wenhao Liu, Muling Wu, Shihan Dou, Changze Lv, Xiaohua Wang, Xiaoqing Zheng, and Xuanjing Huang. Revisiting jailbreaking for large language models: A representation engineering perspective. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3158–3178, Abu Dhabi, UAE, January 2025c. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.212/.

Jingjing Liu, Yonghui Wu, Hao Zhou, Qiying Yu, Chengyi Wang, Zhiqi Lin, Chi Zhang, Jiangjie Chen, Ya-Qin Zhang, Zheng Zhang, Xin Liu, Yuxuan Tong, Mingxuan Wang, Xiangpeng Wei, Lin Yan, Yuxuan Song, Wei-Ying Ma, Yu Yue, Mu Qiao, Haibin Lin, Mofan Zhang, Jinhua Zhu, Guangming Sheng, Wang Zhang, Weinan Dai, Hang Zhu, Gaohong Liu, Yufeng Yuan, Jiaze Chen, Bole Ma, Ruofei Zhu, Tiantian Fan, Xiaochen Zuo, Lingjun Liu, and Hongli Yu. Dapo: An open-source llm reinforcement learning system at scale, 2025a. URL https://arxiv.org/abs/2503.14476.

Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Zhu JianHao, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. Aligning large language models with human preferences through representation engineering. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10619–10638, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.572. URL https://aclanthology.org/2024.acl-long.572/.

Yang Liu, Haiyang Yu, Fei Huang, Yongbin Li, Rongwu Xu, Kun Wang, Zhenhong Zhou, Xinghua Zhang, and Junfeng Fang. On the role of attention heads in large language model safety, 2025b. URL https://arxiv.org/abs/2410.13708.

Yifan Luo, Zhennan Zhou, Meitan Wang, and Bin Dong. Jailbreak instruction-tuned llms via end-of-sentence mlp re-weighting. *arXiv preprint arXiv:2410.10150*, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding an attention head. *arXiv preprint arXiv:2310.04625*, 2023.

Julian Minder, Clément Dumas, Bilal Chughtai, and Neel Nanda. Robustly identifying concepts introduced during chat fine-tuning using crosscoders. In *Sparsity in LLMs (SLLM): Deep Dive into Mixture of Experts, Quantization, Hardware, and Inference*, 2025.

Neel Nanda, Max Tegmark, Senthooran Rajamanoharan, Joshua Engels, and Subhash Kantamneni. Are sparse autoencoders useful? a case study in sparse probing, 2025. URL https://arxiv.org/abs/2502.16681.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Domenic Rosati, Jan Wehner, Kai Williams, Lukasz Bartoszcze, Robie Gonzales, carsten maple, Subhabrata Majumdar, Hassan Sajjad, and Frank Rudzicz. Representation noising: A defence mechanism against harmful finetuning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=eP9auEJqFg.

Cody Rushing and Neel Nanda. Explorations of self-repair in language models. *arXiv preprint arXiv:2402.15390*, 2024.

Mahdi Sabbaghi, Paul Kassianik, George Pappas, Yaron Singer, Amin Karbasi, and Hamed Hassani. Adversarial reasoning at jailbreaking time. *arXiv preprint arXiv:2502.01633*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guobin Shen, Dongcheng Zhao, Yiting Dong, Xiang He, and Yi Zeng. Jailbreak antidote: Runtime safety-utility balance via sparse representation adjustment in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=s20W12XTF8.

Alessandro Stolfo, Ben Peng Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. Confidence regulation neurons in language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=0og7nmvDbe.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wozhdnRCtw.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Cheng Wang, Yue Liu, Baolong Bi, Duzhen Zhang, Zhong-Zhi Li, Yingwei Ma, Yufei He, Shengju Yu, Xinfeng Li, Junfeng Fang, et al. Safety in large reasoning models: A survey. *arXiv preprint arXiv:2504.17704*, 2025.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 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. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=K6xxnKN2gm.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. *arXiv preprint arXiv:2404.15574*, 2024.

Zhihao Xu, Ruixuan Huang, Changyu Chen, and Xiting Wang. Uncovering safety risks of large language models through concept activation vector. *Advances in Neural Information Processing Systems*, 37:116743–116782, 2024.

Hanqi Yan, Hainiu Xu, and Yulan He. Thinking hard, going misaligned: Emergent misalignment in llms. *arXiv preprint arXiv:2509.00544*, 2025.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Xin Yi, Shunfan Zheng, Linlin Wang, Gerard de Melo, Xiaoling Wang, and Liang He. Nlsr: Neuron-level safety realignment of large language models against harmful fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25706–25714, 2025.

Kayo Yin and Jacob Steinhardt. Which attention heads matter for in-context learning? In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=C7XmEByCFv.

Qingyu Yin, Chak Tou Leong, Hongbo Zhang, Minjun Zhu, Hanqi Yan, Qiang Zhang, Yulan He, Wenjie Li, Jun Wang, Yue Zhang, and Linyi Yang. Constrain alignment with sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=BCKSxOFX85.

Bowen Yu, Xiong-Hui Chen, Junyang Lin, Jingren Zhou, Chujie Zheng, An Yang, Rui Men, Chang Gao, Mingze Li, Kai Dang, Yuqiong Liu, and Shixuan Liu. Group sequence policy optimization, 2025. URL https://arxiv.org/abs/2507.18071.

Wojciech Zaremba, Evgenia Nitishinskaya, Boaz Barak, Stephanie Lin, Sam Toyer, Yaodong Yu, Rachel Dias, Eric Wallace, Kai Xiao, Johannes Heidecke, et al. Trading inference-time compute for adversarial robustness. *arXiv preprint arXiv:2501.18841*, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*, 2023.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*, 2024.

Yichi Zhang, Zihao Zeng, Dongbai Li, Yao Huang, Zhijie Deng, and Yinpeng Dong. Realsafe-r1: Safety-aligned deepseek-r1 without compromising reasoning capability. *arXiv preprint arXiv:2504.10081*, 2025.

Yiran Zhao, Wenxuan Zhang, Yuxi Xie, Anirudh Goyal, Kenji Kawaguchi, and Michael Shieh. Understanding and enhancing safety mechanisms of LLMs via safety-specific neuron. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yR47RmND1m.

Baihui Zheng, Boren Zheng, Kerui Cao, Yingshui Tan, Zhendong Liu, Weixun Wang, Jiaheng Liu, Jian Yang, Wenbo Su, Xiaoyong Zhu, et al. Beyond safe answers: A benchmark for evaluating true risk awareness in large reasoning models. *arXiv preprint arXiv:2505.19690*, 2025.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Shreedhar Jangam, Jayanth Srinivasa, Gaowen Liu, Dawn Song, and Xin Eric Wang. The hidden risks of large reasoning models: A safety assessment of r1. *arXiv preprint arXiv:2502.12659*, 2025a.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. On the role of attention heads in large language model safety. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=h0Ak8A5yqw.

Junda Zhu, Lingyong Yan, Shuaiqiang Wang, Dawei Yin, and Lei Sha. Reasoning-to-defend: Safety-aware reasoning can defend large language models from jailbreaking. *arXiv preprint arXiv:2502.12970*, 2025.

Minjun Zhu, Linyi Yang, Yifan Wei, Ningyu Zhang, and Yue Zhang. Locking down the finetuned llms safety. *arXiv preprint arXiv:2410.10343*, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023a.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023b.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, J Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=IbIB8SBKFV.

USE OF LLMS

In the preparation of this manuscript, we utilized LLMs as a writing assistant. The use of LLMs was confined to tasks such as improving grammar, refining phrasing for clarity, and polishing the overall language of the paper. All core scientific contributions, including the conceptualization of ideas, the design and execution of experiments, the analysis of results, and the conclusions, are entirely the work of the human authors. The authors bear full responsibility for the content and claims presented in this work.

A PROBER

In this section, we provide a detailed description of the architecture, data collection, and training procedure for the refusal prober used in our experiments. This prober is a linear classifier designed to predict whether a model will refuse a harmful request based on its internal hidden states.

Prober Architecture. The prober is implemented as a simple linear classifier. Given a hidden state vector $\boldsymbol{h} \in \mathbb{R}^d$ from the reasoning model, where d is the hidden dimension size, the prober computes a single logit. This is followed by a sigmoid function to produce a refusal probability, as defined in Equation 2. The model is implemented in PyTorch using a single ‘torch.nn.Linear‘ layer. We use the Binary Cross-Entropy with Logits loss function (‘nn.BCEWithLogitsLoss‘) for training, which is numerically stable and suitable for binary classification tasks.

Dataset Collection and Preprocessing. To train the prober, we constructed a balanced dataset of hidden states corresponding to both refusal and non-refusal responses.

- **Refusal Examples (Positive Class):** We collected examples where the model refused to comply with a harmful prompt. These were sourced from the AdvBench dataset (Zou et al., 2023b). An output was labeled as a refusal if it contained keywords like "I'm sorry," "I cannot," or similar phrases within the first 32 tokens of the response.
- **Non-Refusal Examples (Negative Class):** For the non-refusal class, we used harmless prompts and their corresponding compliant answers from the UltraChat-SFT dataset (Ding et al., 2023).

For each example in both classes, we fed the full input sequence (user prompt + model's chain of thought + thinking-end template) into the target reasoning model. We then extracted the hidden state vector from the **final token position** at the **last transformer layer**. These hidden state vectors form the training data for our prober.

Training Details. The prober was trained on the collected hidden states. Before training, we balanced the dataset by randomly downsampling the larger class to match the number of samples in the smaller class, ensuring an equal number of refusal and non-refusal examples. The full dataset was then split into training (80%) and validation (20%) sets.

The training hyperparameters are as follows:

- **Optimizer:** Adam
- **Learning Rate:** 1×10^{-3}
- **Batch Size:** 256
- **Epochs:** 5

We selected the model checkpoint that achieved the highest accuracy on the validation set. As reported in Section 3, the final prober achieved over 95% validation accuracy on in-distribution data and demonstrated strong generalization to an out-of-distribution (OOD) dataset, JailbreakBench. This high accuracy confirms that the prober reliably captures the model's refusal intention from its final hidden state.

B Supervised Fine-tuning Details

We performed full-parameter supervised fine-tuning (SFT) to repair the safety alignment of the reasoning models using the data subsets selected by our Cliff-as-a-Judge method. The entire training process was conducted using the LLaMA-Factory library. The base model for the fine-tuning experiments reported in Section 5 was `deepseek-ai/DeepSeek-R1-Distill-Qwen-7B`. We utilized DeepSpeed ZeRO Stage 2 for efficient distributed training. The key hyperparameters and configuration settings are detailed below:

- **Finetuning Type:** Full-parameter SFT
- **Learning Rate:** 5×10^{-6}
- **LR Scheduler:** Linear
- **Epochs:** 1.0
- **Batch Size:** 1 per device with 4 gradient accumulation steps, resulting in an effective batch size of 4.
- **Optimizer:** AdamW (`adamw_torch`)
- **Precision:** BF16
- **Max Sequence Length:** 16,384
- **Attention Implementation:** Flash Attention
- **Prompt Template:** `deepseekr1`
- **Distributed Training:** DeepSpeed ZeRO Stage 2