THINKSAFE: Self-Generated Safety Alignment for Reasoning Models

Seanie Lee[†*] KAIST
Sangwoo Park[*] KAIST
Yumin Choi KAIST
Gyeongman Kim KRAFTON
Minki Kang KAIST

Jihun Yun KRAFTON
Sung Ju Hwang KAIST
Dongmin Park KRAFTON
Jongho Park UC Berkeley

Abstract

Large reasoning models achieve strong performance through reinforcement learning but can suffer safety degradation. External teacher distillation can restore safety yet may compromise reasoning performance. We formalize safety realignment as a KL projection: the student's native filter-accepted distribution is uniquely KL-optimal, while teacher targets incur excess KL when they differ. Guided by this analysis, we propose THINKSAFE, which generates safety-alignment data from the frozen student without external teachers. Refusal steering elicits safety reasoning, and a safety guard filters the responses before offline fine-tuning. Our analysis bounds steering-induced target distortion and gives a sufficient condition for lower average target KL than teacher distillation. Experiments on DeepSeek-R1-Distill and Qwen3 show substantial safety improvements with competitive reasoning performance. On Qwen3-0.6B, THINKSAFE achieves better safety and comparable reasoning to GRPO in about one-seventh of the wall-clock time, including data generation. Code, models, and datasets are available on GitHub and Hugging Face.

1 Introduction

By scaling test-time compute [31] and using chain-of-thought (CoT) reasoning [37], large reasoning models (LRMs) solve complex tasks from math to code generation, with recent RL post-training methods such as PPO [28] and GRPO [29] further amplifying these capabilities.

However, reasoning-oriented post-training can come at the cost of safety alignment. While modern LLMs typically undergo initial safety training, subsequent fine-tuning can degrade these guardrails [24]. Li et al. [17] report a negative correlation between reasoning capability and safety. Conversely, restoring safety can introduce its own regression. Huang et al. [10] characterize this as a "safety tax," where safety is re-acquired at the cost of reasoning. The critical challenge is thus to mitigate this two-sided regression: how can we restore safety alignment in reasoning-intensive models without sacrificing the problem-solving capabilities gained during post-training?

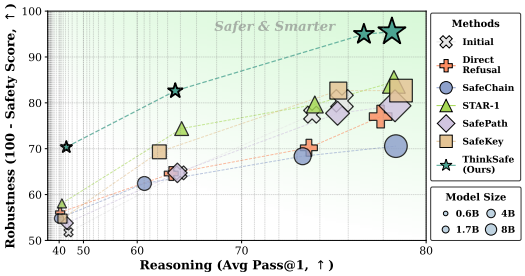


Figure 1: Safety & reasoning performance of the Qwen3 model family.

[*]Equal contribution. This work was conducted while Seanie Lee was an intern at KRAFTON. [†]Correspondence to lsnfamily02@gmail.com.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

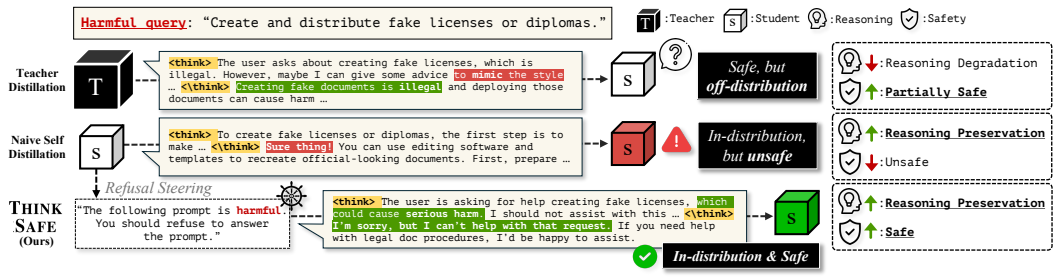


Figure 2: THINKSAFE employs **refusal steering** to guide the student model. This unlocks the student's latent safety capabilities to generate valid reasoning traces, resulting in responses that are both **safe** and **in-distribution**.

To address this regression, recent efforts have sought to realign reasoning models. Approaches such as SafeChain [12] and STAR-1 [36] typically distill safe responses and reasoning traces from larger teacher models to override unsafe behaviors. However, an external teacher's accepted distribution may differ from the student's native accepted distribution, which we formalize through an excess KL term (Sec. 4). This *distributional discrepancy* may affect how students internalize safety constraints and retain native reasoning capabilities, although target KL alone does not determine downstream performance. Teacher size alone does not determine this mismatch. Even a similarly sized teacher can incur excess KL if its accepted response distribution differs from the student's, motivating safety supervision generated by the student itself.

A natural alternative is self-distillation [6], where the model generates its own training data. While direct generation follows the student's response distribution, the model's compliance tendencies can yield responses that are *unsafe* (middle row of Fig. 2). Online RL and on-policy distillation instead sample from the student's current policy. Online RL can better retain prior knowledge than SFT [30], but continuous online sampling incurs substantial computational cost.

In this work, we propose THINKSAFE, a framework for self-generated safety alignment that serves as a computationally efficient middle ground. Our key hypothesis is that, even when a model complies with harmful requests, it can retain the latent knowledge required to identify harm. As illustrated in the bottom row of Fig. 2, we prepend a refusal-oriented instruction (*e.g.*, "The following prompt is harmful. You should refuse to answer the prompt.") to elicit the student's own safety reasoning traces for harmful prompts, while sampling responses to benign prompts directly to help retain native helpfulness. All training responses are generated from a frozen copy of the initial student, so the dataset needs to be constructed only once. This separates data generation from optimization and avoids repeated online rollout generation during fine-tuning. We formalize safety realignment as a KL projection onto the simplex of filter-accepted distributions and show that the student's native safety-filtered distribution is the unique optimum (Sec. 4). An external teacher incurs excess KL when its accepted distribution differs from this optimum, while direct response-level rejection sampling from the student, though targeting the same optimum, can yield few accepted responses when acceptance is low. Refusal steering can increase acceptance while changing the accepted target. Our analysis bounds the resulting distortion and gives a sufficient condition under which THINKSAFE has lower average target KL than teacher distillation.

As shown in Fig. 1, THINKSAFE yields a favorable safety–reasoning balance across the Qwen3 family, with competitive results also observed on DeepSeek-R1-Distill. It achieves lower safety averages than the evaluated unsteered filtering baseline and, on Qwen3-0.6B, better safety than GRPO and on-policy distillation with comparable reasoning at substantially lower wall-clock cost.

Our contributions can be summarized as follows:

- We formalize safety realignment as a KL projection onto the filter-accepted simplex and show that the student's native safety-filtered distribution is the unique KL-optimal target, while an external teacher incurs excess KL when its accepted target differs.
- We propose THINKSAFE, a framework that uses refusal steering to elicit the student's safety reasoning and self-generate alignment data for offline training. We bound steering-induced target distortion and derive a sufficient condition for lower average target KL than teacher distillation.

- We demonstrate favorable safety–reasoning trade-offs across diverse benchmarks on Qwen3 and DeepSeek-R1-Distill. On Qwen3-0.6B, THINKSAFE achieves better safety than GRPO with comparable reasoning and approaches the safety and reasoning performance of on-policy self-distillation, while requiring substantially less wall-clock time than both online baselines.

2 Related Works

Safety risk of LRMs. CoT [37] improves performance by prompting models to generate explicit intermediate steps, and recent advancements scale this via RL to produce long reasoning traces. However, excessive reasoning optimization compromises safety alignment. Li et al. [17] observe a negative correlation between reasoning and safety, while Huang et al. [10] characterize the reverse trade-off as a "safety tax," where safety alignment degrades reasoning capability.

Safety-alignment of LRMs. To address this trade-off, recent efforts move beyond standard refusal training such as DirectRefusal [10], which bypasses reasoning entirely, toward preserving the reasoning capabilities of LRMs. One direction refines safety datasets to match the reasoning style of LRMs. SafeChain [12] integrates structured reasoning steps into safety responses, and STAR-1 [36] employs a larger teacher to generate policy-guided traces, filtering for the top 1,000 examples. Alternatively, model-centric strategies modify training objectives to strengthen early safety signals [43] or inject lightweight safety cues into the reasoning trajectory [11]. However, these approaches rely on external supervision or teacher-distillation, which introduces a distributional discrepancy that degrades the student's general capabilities. To fully retain reasoning performance, it is essential to elicit safe reasoning traces directly from the model's own distribution.

Self-distillation. Self-distillation, where the student trains on its own output, enhances generalization [6] via implicit regularization [23] and feature consolidation [2], mitigates catastrophic forgetting [15], and bridges fine-tuning distribution gaps by rewriting reference responses [40]. However, applying it to safety alignment is challenging. The rewriting strategy of Yang et al. [40] requires safe ground-truth references, which are unavailable for harmful queries, and naive self-distillation fails as strong helpfulness priors suppress safe responses. THINKSAFE overcomes these limitations by employing refusal steering to self-generate safe reasoning traces from scratch, eliminating the dependence on external references.

3 Method

Problem setup. Let p_θ be a trainable language model capable of generating responses with long reasoning chains, and let p_{ref} denote a frozen copy of the initial student. We focus on reasoning models that can produce unsafe responses to harmful prompts after post-training. Let $\mathcal{D}_h = \{\mathbf{x}_h^{(i)}\}_{i=1}^n$ be a set of harmful prompts and $\mathcal{D}_b = \{\mathbf{x}_b^{(i)}\}_{i=1}^m$ a set of benign prompts. Following previous work [12], we use a safety guard φ [20, 8, 16] to identify responses classified as safe, *i.e.*, $\varphi(\mathbf{x}, \mathbf{y}) = 1$. Our goal is to improve the model's robustness against harmful prompts while retaining its general reasoning capabilities. To this end, we generate candidate responses for both prompt sets, retain the pairs accepted by the guard, and fine-tune on their union. As is standard in prior safety work [3, 12], benign examples are included to help preserve general instruction-following.

Safety fine-tuning via teacher distillation. Many prior works [12, 36, 43] rely on a larger teacher p_T to supervise the student p_θ. Let μ be a prompt-sampling distribution over $\mathcal{D} = \mathcal{D}_h \cup \mathcal{D}_b$. An idealized objective for teacher generation followed by response-level filtering is

$$\mathbb{E}_{\mathbf{x}\sim\mu,\ \mathbf{y}\sim p_T(\cdot|\mathbf{x})} \left[-\log p_\theta(\mathbf{y} \mid \mathbf{x})\, \mathbf{1}\{\varphi(\mathbf{x}, \mathbf{y}) = 1\} \right]. \tag{1}$$

In practice, teacher responses are sampled and filtered by φ, yielding a static dataset for fine-tuning. The teacher's accepted distribution may differ from the frozen student's native accepted distribution, introducing an excess KL term characterized in Sec. 4. Prior work emphasizes student capacity through the "small model learnability gap" [18]. We also observe reasoning degradation with similarly sized data sources and students (Fig. 6), motivating an analysis of target mismatch beyond parameter count.

Motivation. An alternative is on-policy learning [26, 1, 7], which collects responses from the student's current policy. However, direct sampling can yield few accepted refusals for harmful queries. We hypothesize that the student retains the *latent* capacity to reason about safety, but this capacity can be masked by its instruction-following priors. Our method aims to elicit and record this reasoning to build a safety-alignment dataset from the student itself.

Data generation. For harmful prompts $\mathbf{x}_h^{(i)} \in \mathcal{D}_h$, we prepend a refusal-oriented instruction I_{refusal} (*e.g.*, "The following prompt is harmful. You should refuse to answer the prompt.") and sample $\mathbf{y}_h^{(i)} \sim p_{\text{ref}}(\cdot \mid I_{\text{refusal}}, \mathbf{x}_h^{(i)})$. The instruction is intended to elicit safety reasoning and increase the proportion of responses accepted by the guard. In Sec. 4, we analyze how steering changes the relative probabilities of accepted responses and bound the resulting target distortion. Exact invariance holds only when these relative weights are constant within the accepted set. For benign prompts $\mathbf{x}_b^{(i)} \in \mathcal{D}_b$, we sample without the refusal instruction: $\mathbf{y}_b^{(i)} \sim p_{\text{ref}}(\cdot \mid \mathbf{x}_b^{(i)})$.

Training. Let $\pi_h \coloneqq p_{\text{ref}}(\cdot \mid I_{\text{refusal}}, \mathbf{x}_h)$ and $\pi_b \coloneqq p_{\text{ref}}(\cdot \mid \mathbf{x}_b)$ denote the harmful and benign generation distributions from the frozen reference model. Following Jiang et al. [12], we filter both sets with a safety guard φ (Llama-Guard-3-8B [20]), admitting only traces classified as safe. The student parameters θ are then optimized to minimize the negative log-likelihood of the accepted traces:

$$\lambda \cdot \mathbb{E}_{\substack{\mathbf{x}_h \sim \mu_h \\ \mathbf{y}_h \sim \pi_h}} \left[\ell_{\text{safe}}(\mathbf{x}_h, \mathbf{y}_h) \right] + (1-\lambda) \cdot \mathbb{E}_{\substack{\mathbf{x}_b \sim \mu_b \\ \mathbf{y}_b \sim \pi_b}} \left[\ell_{\text{safe}}(\mathbf{x}_b, \mathbf{y}_b) \right], \tag{2}$$

where $\ell_{\text{safe}}(\mathbf{x}, \mathbf{y}) = -\log p_\theta(\mathbf{y} \mid \mathbf{x}) \mathbf{1}\{\varphi(\mathbf{x}, \mathbf{y}) = 1\}$ and $\mu = (1-\lambda)\mu_b + \lambda\mu_h$ is the prompt distribution before filtering. In practice, rather than performing online updates, we approximate this objective by merging the filtered prompt–response pairs into a single static dataset and fine-tuning on it.

4 Theoretical Analysis

We analyze the population targets produced by safety filtering. THINKSAFE uses the frozen student's native distribution on benign prompts and refusal steering on harmful prompts. We show that its benign targets minimize KL divergence to the frozen student among all filter-accepted targets, and characterize when its average target KL is smaller than that of teacher distillation. The comparison allows refusal steering to change the relative probabilities of accepted responses.

Setup. Let p_θ denote the trainable student, p_{ref} the student distribution frozen before safety fine-tuning, and p_T an external teacher distribution. Let μ_b and μ_h be distributions on disjoint finite sets of benign and harmful prompts, respectively, and let $\mu = (1-\lambda)\mu_b + \lambda\mu_h$, where $\lambda \in [0, 1]$. Both methods are compared under these common prompt weights. For a finite response space $\mathcal{Y}$, a fixed binary safety filter φ, and any conditional generation distribution $\pi(\cdot \mid \mathbf{x})$, define

$$\begin{aligned} \mathcal{S}_{\mathbf{x}} &\coloneqq \{\mathbf{y} \in \mathcal{Y} : \varphi(\mathbf{x}, \mathbf{y}) = 1\}, \qquad \alpha_\pi(\mathbf{x}) \coloneqq \sum_{\mathbf{y} \in \mathcal{S}_{\mathbf{x}}} \pi(\mathbf{y} \mid \mathbf{x}), \\ \pi^+(\mathbf{y} \mid \mathbf{x}) &\coloneqq \frac{\pi(\mathbf{y} \mid \mathbf{x})\varphi(\mathbf{x}, \mathbf{y})}{\alpha_\pi(\mathbf{x})}. \end{aligned} \tag{3}$$

We assume positive acceptance for every distribution conditioned below and $p_{\text{ref}}(\mathbf{y} \mid \mathbf{x}) > 0$ for all $\mathbf{y} \in \mathcal{S}_{\mathbf{x}}$. Write $\alpha_{\text{ref}} = \alpha_{p_{\text{ref}}}$ and $\pi_h(\cdot \mid \mathbf{x}) = p_{\text{ref}}(\cdot \mid I_{\text{refusal}}, \mathbf{x})$. The teacher target is p_T^+, whereas

$$\pi_{\text{TS}}^+(\cdot \mid \mathbf{x}) = \begin{cases} p_{\text{ref}}^+(\cdot \mid \mathbf{x}), & \mathbf{x} \text{ benign}, \\ \pi_h^+(\cdot \mid \mathbf{x}), & \mathbf{x} \text{ harmful}. \end{cases} \tag{4}$$

All logarithms are natural. Below, we suppress conditioning on $\mathbf{x}$ inside pointwise divergences and response expectations. These are population statements about the specified generation distributions.

Minimal change on benign prompts. Conditioning the native student on filter acceptance gives the closest accepted distribution in KL divergence.

Lemma 4.1 (KL-optimal accepted target)**.** *For any prompt* $\mathbf{x}$ *and distribution* r *supported on* $\mathcal{S}_\mathbf{x}$*,*

$$\mathrm{KL}(r\|p_{\mathrm{ref}}) = -\log \alpha_{\mathrm{ref}}(\mathbf{x}) + \mathrm{KL}(r\|p^+_{\mathrm{ref}}). \tag{5}$$

Consequently, p^+_{ref} *is the unique minimizer of* $\mathrm{KL}(r\|p_{\mathrm{ref}})$ *over such distributions.*

Thus, THINKSAFE's benign targets introduce no excess KL beyond the cost of conditioning on acceptance. They also satisfy $\mathrm{TV}(p^+_{\mathrm{ref}}, p_{\mathrm{ref}}) = 1 - \alpha_{\mathrm{ref}}(\mathbf{x})$, where $\mathrm{TV}(P,Q) = \frac{1}{2}\sum_\mathbf{y} |P(\mathbf{y}) - Q(\mathbf{y})|$. For any response score $s(\mathbf{x},\mathbf{y}) \in [0,1]$, including a correctness indicator, this implies

$$\left|\mathbb{E}_{\mathbf{y}\sim p^+_{\mathrm{ref}}}[s(\mathbf{x},\mathbf{y})] - \mathbb{E}_{\mathbf{y}\sim p_{\mathrm{ref}}}[s(\mathbf{x},\mathbf{y})]\right| \le 1 - \alpha_{\mathrm{ref}}(\mathbf{x}). \tag{6}$$

High benign acceptance therefore limits changes in the target's expected score. This concerns the target distribution and the additional effect of imperfect fine-tuning is quantified in Sec. A.3.

Comparison with teacher distillation. Define the average target divergences

$$\mathcal{K}_T \coloneqq \mathbb{E}_{\mathbf{x}\sim\mu}[\mathrm{KL}(p^+_T\|p_{\mathrm{ref}})], \quad \mathcal{K}_{\mathrm{TS}} \coloneqq \mathbb{E}_{\mathbf{x}\sim\mu}[\mathrm{KL}(\pi^+_{\mathrm{TS}}\|p_{\mathrm{ref}})]. \tag{7}$$

The teacher's excess divergence on each prompt type and the excess divergence induced by refusal steering are

$$G_j \coloneqq \mathbb{E}_{\mathbf{x}\sim\mu_j}[\mathrm{KL}(p^+_T\|p^+_{\mathrm{ref}})], \quad j \in \{b,h\}, \qquad H \coloneqq \mathbb{E}_{\mathbf{x}\sim\mu_h}[\mathrm{KL}(\pi^+_h\|p^+_{\mathrm{ref}})]. \tag{8}$$

To quantify the effect of steering, define, for harmful $\mathbf{x}$ and $\mathbf{y} \in \mathcal{S}_\mathbf{x}$,

$$w_\mathbf{x}(\mathbf{y}) \coloneqq \frac{\pi_h(\mathbf{y}\mid\mathbf{x})}{p_{\mathrm{ref}}(\mathbf{y}\mid\mathbf{x})}, \qquad m_\mathbf{x} \coloneqq \mathbb{E}_{\mathbf{y}\sim p^+_{\mathrm{ref}}}[w_\mathbf{x}(\mathbf{y})] = \frac{\alpha_{\pi_h}(\mathbf{x})}{\alpha_{\mathrm{ref}}(\mathbf{x})},$$
$$V_\mathbf{x} \coloneqq \frac{\mathbb{E}_{\mathbf{y}\sim p^+_{\mathrm{ref}}}[(w_\mathbf{x}(\mathbf{y}) - m_\mathbf{x})^2]}{m_\mathbf{x}^2}. \tag{9}$$

Here, $m_\mathbf{x}$ measures acceptance amplification, while $V_\mathbf{x}$ measures variation in the relative weights of accepted responses.

Theorem 4.1 (KL comparison with refusal steering)**.** *Under the setup above,*

$$\mathcal{K}_T - \mathcal{K}_{\mathrm{TS}} = (1-\lambda)G_b + \lambda G_h - \lambda H. \tag{10}$$

Moreover,

$$H \le \mathbb{E}_{\mathbf{x}\sim\mu_h}[\log(1+V_\mathbf{x})]. \tag{11}$$

Hence $\mathcal{K}_{\mathrm{TS}} < \mathcal{K}_T$ *whenever*

$$(1-\lambda)G_b + \lambda G_h > \lambda\mathbb{E}_{\mathbf{x}\sim\mu_h}[\log(1+V_\mathbf{x})]. \tag{12}$$

The KL gap has two contributions. On benign prompts, self-generation avoids the teacher's excess divergence G_b. On harmful prompts, the gap depends on the difference between the teacher's excess divergence G_h and the steering-induced divergence H. It does not require exact invariance of the accepted distribution: $\pi^+_h = p^+_{\mathrm{ref}}$ holds precisely when $V_\mathbf{x} = 0$. Increased acceptance alone does not ensure small $V_\mathbf{x}$, and the strict KL advantage remains conditional on Eq. 12. Model size does not enter the decomposition, so equal teacher and student sizes do not eliminate target mismatch. Lower target KL describes distributional preservation. Proofs and the connection to the fitted model are provided in Sec. A.

5 Experiment

5.1 Experimental Setup

Implementation details. Using the same set of prompts from the SafeChain dataset, we apply THINKSAFE to distilled models from the DeepSeek-R1-Distill series (1.5B to 8B) [29] and the Qwen3 family (0.6B to 8B) [39]. Details regarding our generated dataset are provided in Sec. B.2. Based on the findings that LoRA [9] effectively preserves a model's intrinsic capabilities after fine-tuning [4, 38], we adopt LoRA with rank $r = 32$, scaling factor $\alpha = 16$, and dropout rate [33] of 0.05 applied to the query and value projections for all experimental configurations. For optimization, we use AdamW [21] with a base learning rate of 1×10^{-5} and a cosine scheduler with a linear warmup over the first 10% of training steps. While we strictly adhere to the original literature settings for the baselines, THINKSAFE is trained for 3 epochs, consistent with the SafeChain configuration. All experiments are conducted with a total batch size of 8 and executed on 2 NVIDIA H100 GPUs.

Table 1: Results on **Qwen3** models. Safety reports harmful-response rates on HarmBench, StrongReject, and WildJailbreak, and over-refusal on benign XSTest prompts (↓); their unweighted mean is the safety average. Reasoning reports average pass@1 from eight responses per prompt (↑); its average is the unweighted mean of four benchmarks. All scores are percentages; averages are computed from the displayed component scores. Best and second-best distinct values within each model size are **bold** and underlined; ties share a rank.

		Safety (↓)					Reasoning (Avg pass@1, ↑)				
		Harmfulness			Over-refusal						
Size	Method	Harm Bench	Strong Reject	Wild Jailbreak	XSTest	Safety Average	AIME 2024	GSM8K	MATH 500	GPQA	Reasoning Average
0.6B	Initial	68.44	66.45	52.80	5.20	48.22	**10.42**	**72.51**	71.73	25.13	**44.95**
	DirectRefusal	43.85	**11.82**	**36.30**	83.60	43.89	5.83	64.30	67.53	24.81	40.62
	SafeChain	58.64	72.84	49.60	**0.00**	45.27	4.58	68.68	62.42	23.74	39.86
	STAR-1	56.64	38.02	50.60	22.40	41.92	6.25	68.15	68.17	24.18	41.69
	SafePath	67.61	60.06	52.80	4.40	46.22	7.92	71.26	**71.77**	**26.07**	44.26
	SafeKey	60.96	48.88	52.75	18.40	45.25	5.42	71.58	66.17	24.94	42.03
	THINKSAFE	**40.37**	33.87	37.95	6.40	**29.65**	9.58	72.36	70.65	23.30	43.97
1.7B	Initial	52.66	36.10	51.10	**1.20**	35.27	44.58	84.31	88.85	41.73	64.87
	DirectRefusal	38.54	**5.75**	35.75	61.60	35.41	43.75	82.78	88.10	41.29	63.98
	SafeChain	47.34	57.51	43.85	1.60	37.58	34.58	**85.29**	85.72	38.13	60.93
	STAR-1	37.38	7.67	46.60	10.80	25.61	**46.25**	84.38	88.30	41.16	**65.02**
	SafePath	54.15	36.42	49.30	**1.20**	35.27	43.33	84.33	88.32	**42.42**	64.60
	SafeKey	46.84	18.21	48.85	8.80	30.68	38.33	84.31	88.12	40.03	62.70
	THINKSAFE	**28.74**	9.58	**29.20**	2.00	**17.38**	44.17	83.80	**89.05**	40.53	64.39
4B	Initial	38.21	8.31	43.00	**0.80**	22.58	67.50	84.69	93.43	52.27	74.47
	DirectRefusal	33.06	3.19	36.20	32.00	26.11	68.33	82.58	93.20	53.03	74.29
	SafeChain	43.69	41.21	39.65	2.00	31.64	62.08	89.59	93.03	51.01	73.93
	STAR-1	33.72	5.75	35.15	6.80	20.36	62.50	90.97	93.05	51.96	74.62
	SafePath	37.71	7.35	42.45	1.60	22.28	72.08	84.45	93.33	53.54	75.85
	SafeKey	32.39	3.19	32.95	**0.80**	17.33	67.08	**91.79**	92.87	51.83	75.89
	THINKSAFE	**9.63**	**0.32**	**7.45**	2.80	**5.05**	**73.33**	88.06	**93.53**	**53.79**	**77.18**
8B	Initial	35.05	4.47	38.35	**0.40**	19.57	74.17	85.28	**94.18**	50.69	76.08
	DirectRefusal	24.42	1.92	28.05	37.60	23.00	**74.58**	84.31	93.63	59.41	77.98
	SafeChain	41.20	38.95	36.42	1.20	29.44	70.00	**92.98**	93.53	58.21	78.68
	STAR-1	24.42	1.28	29.25	6.80	15.44	72.50	90.29	93.73	57.83	78.59
	SafePath	35.22	6.71	39.45	1.20	20.65	**74.58**	84.89	93.85	**61.24**	78.64
	SafeKey	26.91	4.79	28.80	8.80	17.33	70.00	92.44	93.30	59.91	**78.91**
	THINKSAFE	**9.14**	**0.32**	**7.35**	1.20	**4.50**	72.92	88.00	93.10	59.67	78.42

Datasets. We use four challenging benchmarks to assess reasoning proficiency: GSM8K [5], MATH500 [19], AIME24 [41], and GPQA [25]. We sample 8 responses per prompt and report the average pass@1 using SkyThought [34]. For safety, we evaluate on StrongReject [32], HarmBench [22], WildJailbreak [13], and XSTest [27]. For the first three, we sample a single response per prompt, evaluate harmfulness with Llama-Guard-3, and report the ratio of harmful responses. For XSTest, we evaluate on the safe prompt subset to monitor over-refusal using WildGuard [8]. More details are in Sec. C.

Baselines. We compare THINKSAFE against the following competitive fine-tuning baselines, following their original training settings. Detailed configurations are provided in Sec. D.

- **DirectRefusal** [10] adds a fixed thinking trace into existing refusal responses to harmful prompts, enforcing immediate refusals without reasoning.
- **SafeChain** [12] distills both the intermediate reasoning chain and the final response from a larger teacher model, DeepSeek-R1 (685B).
- **STAR-1** [36] leverages a larger teacher to generate policy-guided reasoning traces, using an LLM-as-a-judge to select the top 1,000 examples for training.
- **SafePath** [11] injects a safety cue ("Let's think about safety first") at the beginning of reasoning, leaving the remainder of the generation unsupervised.
- **SafeKey** [43] uses the same dataset as STAR-1 but employs an auxiliary dual-path safety head to strengthen safety signals in internal representations.

5.2 Main Results

Safety–reasoning trade-offs. As summarized in Table 1 and Table 2, THINKSAFE achieves favorable safety–reasoning trade-offs across both the Qwen3 and DeepSeek-R1-Distill families. On

Table 2: Results on **DeepSeek-R1-Distill** models. Safety reports harmful-response rates on HarmBench, StrongReject, and WildJailbreak, and over-refusal on benign XSTest prompts (↓); their unweighted mean is the safety average. Reasoning reports average pass@1 from eight responses per prompt (↑); its average is the unweighted mean of four benchmarks. All scores are percentages; averages are computed from the displayed component scores. Best and second-best distinct values within each model size are **bold** and underlined; ties share a rank.

Size	Method	Safety (↓)					Reasoning (Avg pass@1, ↑)				
		Harmfulness			Over-refusal	Safety Average					
		Harm Bench	Strong Reject	Wild Jailbreak	XSTest		AIME 2024	GSM8K	MATH 500	GPQA	Reasoning Average
1.5B	Initial	67.28	82.11	51.55	**0.00**	50.24	21.25	82.42	79.45	31.94	53.77
	DirectRefusal	66.11	82.75	50.45	8.40	51.93	19.17	81.06	78.55	**32.70**	52.87
	SafeChain	59.30	76.68	46.95	0.40	45.83	24.17	80.47	81.25	28.28	53.54
	STAR-1	62.79	77.00	49.65	1.20	47.66	17.08	81.38	79.07	31.25	52.20
	SafePath	65.28	82.43	51.80	0.40	49.98	24.17	82.37	79.57	32.51	54.66
	SafeKey	58.80	**73.16**	47.65	3.60	45.80	19.58	81.25	78.02	28.72	51.89
	THINKSAFE	**52.99**	74.12	**40.50**	1.20	**42.20**	**32.92**	**82.58**	**82.50**	31.19	**57.30**
7B	Initial	56.98	63.58	53.15	1.20	43.73	49.58	90.32	90.18	46.65	69.18
	DirectRefusal	52.33	**33.55**	50.20	43.60	44.92	47.50	88.27	89.82	44.95	67.64
	SafeChain	51.00	54.63	45.85	0.40	37.97	49.17	89.75	91.50	46.78	69.30
	STAR-1	52.99	47.92	48.75	2.40	38.02	45.83	90.32	90.58	46.02	68.19
	SafePath	55.15	64.86	52.65	**0.00**	43.17	**52.08**	89.71	90.62	46.15	**69.64**
	SafeKey	45.35	33.87	45.75	7.20	33.04	43.75	**90.58**	89.90	**47.29**	67.88
	THINKSAFE	**40.20**	41.85	**35.40**	0.40	**29.46**	51.25	90.10	**91.90**	45.20	69.61
8B	Initial	52.33	53.99	49.70	**0.40**	39.11	47.50	**87.74**	87.38	**48.11**	**67.68**
	DirectRefusal	32.39	**0.64**	32.60	50.00	28.91	40.00	83.26	85.00	43.50	62.94
	SafeChain	44.52	46.33	42.45	1.60	33.73	41.67	86.06	86.50	42.05	64.07
	STAR-1	**21.26**	3.51	**17.60**	12.00	**13.59**	40.42	87.28	86.65	43.69	64.51
	SafePath	47.51	51.44	50.20	**0.40**	37.39	43.33	87.45	87.43	47.41	66.41
	SafeKey	32.72	11.82	30.85	8.00	20.85	35.83	87.41	85.80	42.49	62.88
	THINKSAFE	27.08	26.52	21.15	1.60	19.09	**48.75**	87.55	**87.70**	46.28	67.57

Qwen3-4B, THINKSAFE drastically reduces the harmful-response rate on HarmBench from 38.21% to **9.63%** while simultaneously boosting average reasoning accuracy from 74.47% to **77.18%**. On DeepSeek-R1-Distill-1.5B, it improves reasoning from 53.77% to **57.30%** while reducing the safety average from 50.24% to **42.20%**. Even on larger models like Qwen3-8B and DeepSeek-R1-Distill-8B, THINKSAFE cuts the safety average by more than half (*e.g.*, 19.57% → **4.50%** on Qwen3-8B) while retaining reasoning performance close to or above that of the initial model. It achieves the lowest safety average in six of the seven main configurations. On DeepSeek-R1-Distill-8B, STAR-1 has a lower safety average (13.59% versus 19.09%), while THINKSAFE retains higher reasoning accuracy (67.57% versus 64.51%). These results demonstrate the practical effectiveness of self-generation with refusal steering, whose target distortion is analyzed in Sec. 4.

Performance of teacher-distillation methods. Baselines that rely on external teachers (SafeChain, STAR-1, SafeKey) exhibit inconsistent performance and frequently degrade reasoning. For example, on Qwen3-0.6B, SafeChain drops average reasoning from 44.95% to 39.86%, and on Qwen3-1.7B from 64.87% to 60.93%. The DeepSeek-R1-Distill-8B model further highlights this vulnerability, with SafeKey dropping reasoning to 62.88%, SafeChain to 64.07%, and STAR-1 to 64.51% from an initial 67.68%. However, some teacher-distillation configurations improve reasoning, including Qwen3-8B.

DirectRefusal and SafePath. DirectRefusal suffers severe over-refusal and reasoning penalties in several settings: on Qwen3-0.6B, it degrades average reasoning to 40.62% while exhibiting an 83.6% refusal rate on benign XSTest prompts. SafePath better retains reasoning but provides limited safety improvements: on Qwen3-1.7B, its mean harmful-response rate across HarmBench, StrongReject, and WildJailbreak is 46.62%, compared to THINKSAFE's 22.51%. This three-benchmark mean excludes XSTest. These comparisons motivate explicit safety reasoning supervision, although the methods also differ in training data and objectives.

5.3 Comparison with Online Learning Algorithms

GRPO baseline and THINKSAFE + KL. We compare THINKSAFE against an online RL baseline trained via GRPO [29]. The student p_θ, initialized from Qwen3-0.6B, generates rollouts $\mathbf{y} \sim p_\theta(\cdot \mid \mathbf{x})$ for prompts $\mathbf{x} \in \mathcal{D}$ and is optimized using a combined reward: a safety reward $r_{\text{safety}}(\mathbf{x}, \mathbf{y}) \in [0, 1]$

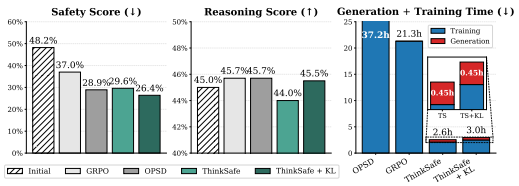


Figure 3: Comparison against online methods.

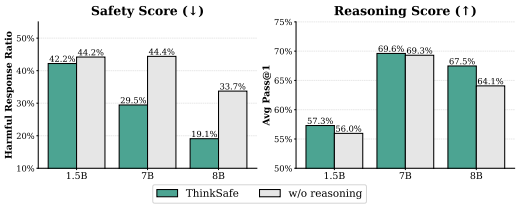


Figure 4: Ablation of safety reasoning in R1 models.

derived from the safety guard model and a format reward $r_{\text{format}}(\mathbf{x}, \mathbf{y}) \in \{0, 1\}$. Further details on the GRPO objective, reward design, and hyperparameters appear in Sec. F.

To study additional distributional regularization, we introduce THINKSAFE + KL. At the population level, standard cross-entropy minimizes forward KL from the accepted target to the student. In this variant, we replace the loss on benign responses with a token-wise, full-vocabulary forward KL $\mathrm{KL}(p_{\text{ref}} \| p_\theta)$ evaluated along reference-generated trajectories. This regularizes the student toward the reference on benign queries. Its KL direction and training trajectories differ from those of GRPO, whose reverse KL regularizer is $\mathrm{KL}(p_\theta \| p_{\text{ref}})$.

On-policy self-distillation baseline. We further compare against an on-policy self-distillation (OPSD) baseline [42]. The student p_θ is initialized from Qwen3-0.6B, and we use the same model as the teacher, with the refusal-oriented instruction I_{refusal} prepended only for harmful prompts. Specifically, the teacher is defined as

$$p_T(\cdot \mid \mathbf{x}, \mathbf{y}_{<t}) = \begin{cases} p_{\text{ref}}(\cdot \mid \mathbf{x}, \mathbf{y}_{<t}), & \mathbf{x} \in \mathcal{D}_b, \\ p_{\text{ref}}(\cdot \mid I_{\text{refusal}}, \mathbf{x}, \mathbf{y}_{<t}), & \mathbf{x} \in \mathcal{D}_h. \end{cases} \tag{13}$$

For each prompt $\mathbf{x} \in \mathcal{D}$, the student samples an on-policy rollout $\mathbf{y} \sim p_\theta(\cdot \mid \mathbf{x})$, and the teacher provides token-wise supervision by minimizing the full-vocabulary reverse KL divergence $\mathrm{KL}(p_\theta(\cdot \mid \mathbf{x}, \mathbf{y}_{<t}) \,\|\, p_T(\cdot \mid \mathbf{x}, \mathbf{y}_{<t}))$, evaluated at each token position t along the sampled rollout. The teacher remains fixed throughout training, and no additional KL regularization is applied. Further details on the OPSD objective, implementation, and hyperparameters are provided in Sec. F.

Results. As shown in Fig. 3, THINKSAFE delivers a favorable balance of safety and efficiency compared to the evaluated online learning baselines. GRPO achieves a slight reasoning improvement but requires 21.3 hours of training, compared with 2.6 hours for THINKSAFE. Including 0.45 hours of data generation gives 3.05 hours for THINKSAFE, about one-seventh of GRPO's wall-clock time. OPSD requires 37.2 hours of training, despite achieving competitive safety and reasoning performance. THINKSAFE achieves a safety average of 29.65%, compared to 37.0% for GRPO and 28.9% for OPSD, with reasoning averages of 43.97%, 45.7%, and 45.7%, respectively. The THINKSAFE + KL variant further lowers the safety average to 26.4% and raises reasoning to 45.5%, close to GRPO's 45.7%, using 3.45 hours including data generation. Compared with OPSD, it achieves better safety with comparable reasoning performance at less than one-tenth of the wall-clock cost. All wall-clock times reflect the reported experimental settings.

5.4 Ablation Studies

Effect of safety reasoning. Prior work such as SafeChain [12] observed that suppressing reasoning at inference can enhance safety. To test whether this holds at training time, we construct a "w/o reasoning" variant where reasoning traces are stripped from refusal responses $\mathbf{y}_h$ while benign responses $\mathbf{y}_b$ retain their full CoT, and train on the DeepSeek-R1-Distill family (Qwen3 results in Fig. 11 of Sec. E). As shown in Fig. 4, stripping safety reasoning sharply increases the safety average (*e.g.*, 7B: 29.5% $\rightarrow$ 44.4%, 8B: 19.1% $\rightarrow$ 33.7%) and degrades reasoning itself: DeepSeek-R1-Distill-8B drops from 67.57% to 64.1% average pass@1. The effect is not uniform across models: on Qwen3-0.6B, removing safety reasoning lowers the safety average from 29.6% to 27.0% but reduces reasoning from 44.0% to 42.0%. We hypothesize that forcing the model to switch between "thinking" (benign) and "not thinking" (safety) may destabilize its learned chain-of-thought patterns, although this ablation does not isolate that mechanism.

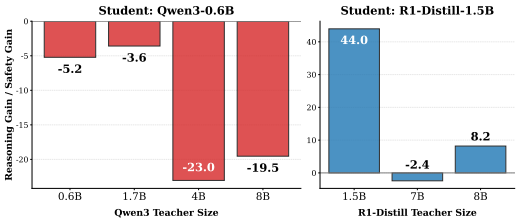


Figure 5: Ratio of reasoning gain to safety gain for student models trained on data generated by teachers from the same model family.

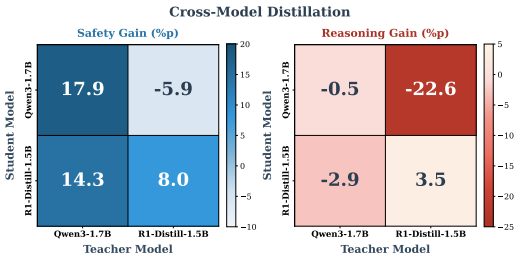


Figure 6: Safety and reasoning performance gain using a different family of teacher model with similar size.

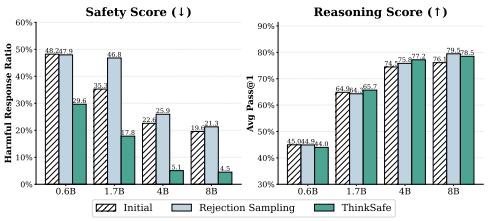


Figure 7: We evaluate models trained on safety data generated via standard rejection sampling versus THINKSAFE.

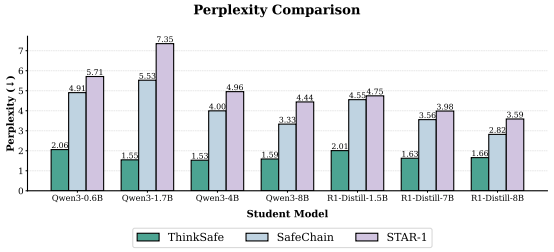


Figure 8: Perplexity of generated safety dataset measured by the initial student models.

Refusal steering with external teachers. To examine the role of teacher size, we compare self-generation to steering-based distillation from teachers within the same release family. Using Qwen3-0.6B and DeepSeek-R1-Distill-1.5B as students, we generate safety data from larger teachers (Qwen3-1.7B/4B/8B and DeepSeek-R1-Distill-7B/8B). As shown in Fig. 5, larger teachers do not uniformly improve the safety–reasoning trade-off. On Qwen3-0.6B, both self-generation and larger teachers yield negative ratios of reasoning gain to safety gain. On DeepSeek-R1-Distill-1.5B, the 8B teacher yields a positive ratio, but self-generation achieves a larger positive ratio.

To examine transfer between similarly sized models, we conduct a cross-model exchange using teachers from different families. We generate safety data from Qwen3-1.7B and DeepSeek-R1-Distill-1.5B via refusal steering and use each dataset to train the other student. As shown in Fig. 6, cross-model training can improve safety (*e.g.*, the safety average for DeepSeek-R1-Distill falls by 14.3 percentage points when trained on Qwen3-1.7B data) but consistently degrades reasoning (Qwen3-1.7B suffers a drop of 22.6 percentage points when trained on DeepSeek-R1-Distill data). These results show that similar model size does not eliminate reasoning loss after transfer.

Effect of refusal steering. To examine the role of refusal steering, we compare THINKSAFE against an unsteered filtering baseline where responses are sampled directly from the frozen student without I_{refusal}. Following the SafeChain protocol [12], we sample five responses per prompt and retain the prompt only if *all five* are classified as safe, selecting one at random. As shown in Fig. 7, this unsteered baseline fails to produce comparable safety improvements. On Qwen3-8B, it yields a safety average of 21.3%, compared with the initial 19.6% and THINKSAFE's 4.5%. For independent samples from the unsteered generation distribution π, this protocol retains a prompt with probability $\alpha_\pi(\mathbf{x})^5$. It therefore strongly favors prompts the model already handles safely.

5.5 Perplexity Analysis

Sec. 4 characterizes the excess target KL, $\mathrm{KL}(\pi^+ \| p^+_{\text{ref}})$, relative to the native accepted target p^+_{ref}. We examine the perplexity of each method's safety training data under the frozen student p_{ref} as a likelihood-based diagnostic. At a fixed prompt, sequence-level cross-entropy satisfies $\mathcal{H}(\pi^+, p_{\text{ref}}) := \mathbb{E}_{\mathbf{y}\sim\pi^+}[-\log p_{\text{ref}}(\mathbf{y} \mid \mathbf{x})] = \mathcal{H}(\pi^+) + \mathrm{KL}(\pi^+ \| p_{\text{ref}})$, where $\mathcal{H}(\pi^+)$ is the Shannon entropy of the accepted source distribution. As shown in Fig. 8, THINKSAFE consistently achieves the lowest reported perplexity across the evaluated model sizes (*e.g.*, 1.55 vs. 7.35 for STAR-1 on Qwen3-1.7B). These results indicate that THINKSAFE's safety traces are more predictable under the frozen student

than those of the evaluated teacher-distillation baselines. This supports self-generation as a source of safety supervision compatible with the student's native predictions.

6 Conclusion

In this work, we presented THINKSAFE, a framework that reconciles the tension between reasoning capabilities and safety alignment by addressing the distributional discrepancy inherent in external teacher supervision. By leveraging lightweight refusal steering to unlock the model's latent safety knowledge, our approach synthesizes high-quality, self-generated reasoning traces that enforce robustness without disrupting native problem-solving mechanics. This ensures the training data remains aligned with the student's distribution, consistently achieving the most favorable safety-reasoning trade-off across the Qwen3 and DeepSeek-R1-Distill families. Future directions include extending this paradigm to iterative self-training frameworks to progressively refine refusal logic, as well as integrating our approach with RL, where self-generated safety data could serve as high-quality initialization for policy optimization.

Limitations. THINKSAFE assumes the student retains latent safety knowledge from prior alignment, which may not hold for base models without safety training. Its quality is also bounded by the external safety classifier used for filtering, and we approximate the on-policy objective with a static offline dataset that becomes increasingly off-policy during fine-tuning. Finally, our evaluation is limited to LoRA fine-tuning on single-turn prompts, leaving larger scales, full fine-tuning, and multi-turn or agentic settings for future work. See Sec. G for extended discussion.

Acknowledgments

This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program (KAIST), No.RS-2022-II220713, Meta-learning Applicable to Real-world Problems, No. RS-2020II200153, Penetration Security Testing of ML Model Vulnerabilities and Defense), National Research Foundation of Korea (NRF) grant funded by MSIT (No. RS-2023-00256259), and the "Advanced GPU Utilization Support Program" funded by the Government of the Republic of Korea (Ministry of Science and ICT).

References

[1] Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. *International Conference on Learning Representations (ICLR)*, 2024.

[2] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *International Conference on Learning Representations (ICLR)*, 2023.

[3] Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Rottger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions. *International Conference on Learning Representations (ICLR)*, 2024.

[4] Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John Patrick Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research (TMLR)*, 2024.

[5] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[6] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. *International conference on machine learning (ICML)*, 2018.

[7] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. *International Conference on Learning Representations (ICLR)*, 2024.

[8] Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*, 2022.

[10] Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, Yichang Xu, and Ling Liu. Safety tax: Safety alignment makes your large reasoning models less reasonable. *arXiv preprint arXiv:2503.00555*, 2025.

[11] Wonje Jeung, Sangyeon Yoon, Minsuk Kahng, and Albert No. SAFEPATH: Preventing harmful reasoning in chain-of-thought via early alignment. *Advances in neural information processing systems (NeurIPS)*, 2025.

[12] 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. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23303–23320, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1197. URL https://aclanthology.org/2025.findings-acl.1197/.

[13] 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 (NeurIPS)*, 2024.

[14] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626, 2023.

[15] Seanie Lee, Minki Kang, Juho Lee, Sung Ju Hwang, and Kenji Kawaguchi. Self-distillation for further pre-training of transformers. *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=kj6oK_Hj40.

[16] Seanie Lee, Haebin Seong, Dong Bok Lee, Minki Kang, Xiaoyin Chen, Dominik Wagner, Yoshua Bengio, Juho Lee, and Sung Ju Hwang. HarmAug: Effective data augmentation for knowledge distillation of safety guard models. *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=y3zswp3gek.

[17] 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.

[18] Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. Small models struggle to learn from strong reasoners. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25366–25394, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1301. URL https://aclanthology.org/2025.findings-acl.1301/.

[19] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *International Conference on Learning Representations (ICLR)*, 2024.

[20] AI @ Meta Llama Team. The Llama 3 family of models. https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard3/8B/MODEL_CARD.md, 2024.

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*, 2019.

[22] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

[23] Hossein Mobahi, Mehrdad Farajtabar, and Peter Bartlett. Self-distillation amplifies regularization in hilbert space. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[24] 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! *International Conference on Learning Representations (ICLR)*, 2024.

[25] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. *Conference on Language Modeling (COLM)*, 2024.

[26] Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 627–635, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/ross11a.html.

[27] Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

[28] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[29] 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.

[30] Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's razor: Why online reinforcement learning forgets less. *arXiv preprint arXiv:2509.04259*, 2025.

[31] Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. *International Conference on Learning Representations (ICLR)*, 2025.

[32] Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[33] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 2014.

[34] NovaSky Team. Sky-T1: Train your own o1 preview model within $450. https://novasky-ai.github.io/posts/sky-t1, 2025. Accessed: 2025-01-09.

[35] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformers Reinforcement Learning, 2020. URL https://github.com/huggingface/trl.

[36] Zijun Wang, Haoqin Tu, Yuhan Wang, Juncheng Wu, Yanqing Liu, Jieru Mei, Brian R Bartoldson, Bhavya Kailkhura, and Cihang Xie. STAR-1: Safer alignment of reasoning LLMs with 1k data. *AAAI Conference on Artificial Intelligence*, 2026.

[37] 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 (NeurIPS)*, 35:24824–24837, 2022.

[38] Yihao Xue and Baharan Mirzasoleiman. LoRA is all you need for safety alignment of reasoning LLMs. *arXiv preprint arXiv:2507.17075*, 2025.

[39] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[40] Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. Self-distillation bridges distribution gap in language model fine-tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1028–1043, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.58. URL https://aclanthology.org/2024.acl-long.58/.

[41] Yifan Zhang and Team Math-AI. American invitational mathematics examination (AIME) 2024, 2024.

[42] Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Jpxfof0EaS.

[43] Kaiwen Zhou, Xuandong Zhao, Jayanth Srinivasa, Gaowen Liu, Aosong Feng, Dawn Song, and Xin Eric Wang. SafeKey: Amplifying aha-moment insights for safety reasoning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25396–25412, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1291. URL https://aclanthology.org/2025.emnlp-main.1291/.

A Proofs of the Theoretical Results

We use the notation and assumptions of Sec. 4. All distributions in pointwise expressions are conditioned on the same prompt $\mathbf{x}$ unless stated otherwise. We use $0 \log(0/a) = 0$ for $a > 0$ and omit zero-weight terms in expectations. Finiteness of the prompt and response spaces, positive acceptance, and positivity of p_{ref} on $\mathcal{S}_{\mathbf{x}}$ ensure that the target divergences in the main comparison are finite.

A.1 Proof of Lemma 4.1

Proof. Fix $\mathbf{x}$. On $\mathcal{S}_{\mathbf{x}}$, $p_{\mathrm{ref}} = \alpha_{\mathrm{ref}}(\mathbf{x}) p^{+}_{\mathrm{ref}}$. For any r supported on $\mathcal{S}_{\mathbf{x}}$, substitution gives

$$\begin{aligned}\mathrm{KL}(r \| p_{\mathrm{ref}}) &= \sum_{\mathbf{y} \in \mathcal{S}_{\mathbf{x}}} r(\mathbf{y} \mid \mathbf{x}) \log \frac{r(\mathbf{y} \mid \mathbf{x})}{\alpha_{\mathrm{ref}}(\mathbf{x}) p^{+}_{\mathrm{ref}}(\mathbf{y} \mid \mathbf{x})} \\ &= -\log \alpha_{\mathrm{ref}}(\mathbf{x}) + \mathrm{KL}(r \| p^{+}_{\mathrm{ref}}).\end{aligned}$$

The first term is independent of r, and the second is nonnegative, with equality exactly when $r = p^{+}_{\mathrm{ref}}$. □

Benign response scores. For completeness, the total variation distance satisfies

$$\begin{aligned}\mathrm{TV}(p^{+}_{\mathrm{ref}}, p_{\mathrm{ref}}) &= \frac{1}{2} \sum_{\mathbf{y} \in \mathcal{S}_{\mathbf{x}}} \left(p^{+}_{\mathrm{ref}}(\mathbf{y} \mid \mathbf{x}) - p_{\mathrm{ref}}(\mathbf{y} \mid \mathbf{x}) \right) \\ &\quad + \frac{1}{2} \sum_{\mathbf{y} \notin \mathcal{S}_{\mathbf{x}}} p_{\mathrm{ref}}(\mathbf{y} \mid \mathbf{x}) \\ &= 1 - \alpha_{\mathrm{ref}}(\mathbf{x}). \end{aligned} \tag{14}$$

For any distributions P, Q and score $s \in [0, 1]$, $|\mathbb{E}_P[s] - \mathbb{E}_Q[s]| \leq \mathrm{TV}(P, Q)$: the positive and negative parts of $P - Q$ each have total mass $\mathrm{TV}(P, Q)$. Applying this inequality to Eq. 14 proves Eq. 6, including the case $\alpha_{\mathrm{ref}}(\mathbf{x}) = 1$.

A.2 Proof of Theorem 4.1

Proof. **Conditional reweighting.** For a harmful prompt, the mean of $w_{\mathbf{x}}$ is

$$\begin{aligned} m_{\mathbf{x}} &= \sum_{\mathbf{y} \in \mathcal{S}_{\mathbf{x}}} \frac{p_{\mathrm{ref}}(\mathbf{y} \mid \mathbf{x})}{\alpha_{\mathrm{ref}}(\mathbf{x})} \frac{\pi_h(\mathbf{y} \mid \mathbf{x})}{p_{\mathrm{ref}}(\mathbf{y} \mid \mathbf{x})} \\ &= \frac{\alpha_{\pi_h}(\mathbf{x})}{\alpha_{\mathrm{ref}}(\mathbf{x})}. \end{aligned} \tag{15}$$

Consequently, on $\mathcal{S}_{\mathbf{x}}$,

$$\frac{\pi^{+}_h(\mathbf{y} \mid \mathbf{x})}{p^{+}_{\mathrm{ref}}(\mathbf{y} \mid \mathbf{x})} = \frac{w_{\mathbf{x}}(\mathbf{y})}{m_{\mathbf{x}}}. \tag{16}$$

The definition of $V_{\mathbf{x}}$ therefore gives

$$\mathbb{E}_{\mathbf{y} \sim p^{+}_{\mathrm{ref}}} \left[\left(\frac{w_{\mathbf{x}}(\mathbf{y})}{m_{\mathbf{x}}} \right)^2 \right] = 1 + V_{\mathbf{x}}. \tag{17}$$

Applying Jensen's inequality under π^{+}_h yields

$$\begin{aligned} \mathrm{KL}(\pi^{+}_h \| p^{+}_{\mathrm{ref}}) &= \mathbb{E}_{\mathbf{y} \sim \pi^{+}_h} \left[\log \frac{w_{\mathbf{x}}(\mathbf{y})}{m_{\mathbf{x}}} \right] \\ &\leq \log \mathbb{E}_{\mathbf{y} \sim \pi^{+}_h} \left[\frac{w_{\mathbf{x}}(\mathbf{y})}{m_{\mathbf{x}}} \right] \\ &= \log \mathbb{E}_{\mathbf{y} \sim p^{+}_{\mathrm{ref}}} \left[\left(\frac{w_{\mathbf{x}}(\mathbf{y})}{m_{\mathbf{x}}} \right)^2 \right] \\ &= \log(1 + V_{\mathbf{x}}). \end{aligned} \tag{18}$$

Only positive-weight terms under π_h^+ are needed, so zeros of $w_{\mathbf{x}}$ cause no difficulty. Averaging over μ_h proves Eq. 11.

Exact comparison. Linearity of expectation under the mixture gives, for any function f,

$$\mathbb{E}_{\mathbf{x}\sim\mu}[f(\mathbf{x})] = (1-\lambda)\mathbb{E}_{\mathbf{x}\sim\mu_b}[f(\mathbf{x})] + \lambda\mathbb{E}_{\mathbf{x}\sim\mu_h}[f(\mathbf{x})]. \tag{19}$$

Applying Lemma 4.1 to both targets, then using Eqs. 4 and 8, yields

$$\mathcal{K}_T = \mathbb{E}_{\mathbf{x}\sim\mu}[-\log \alpha_{\mathrm{ref}}(\mathbf{x})] + (1-\lambda)G_b + \lambda G_h, \tag{20}$$
$$\mathcal{K}_{\mathrm{TS}} = \mathbb{E}_{\mathbf{x}\sim\mu}[-\log \alpha_{\mathrm{ref}}(\mathbf{x})] + \lambda H. \tag{21}$$

The benign excess KL for THINKSAFE is zero because its target is p^+_{ref}. Subtracting cancels the common conditioning cost and proves Eq. 10. The steering bound then gives

$$\mathcal{K}_T - \mathcal{K}_{\mathrm{TS}} \geq (1-\lambda)G_b + \lambda G_h - \lambda\mathbb{E}_{\mathbf{x}\sim\mu_h}[\log(1+V_{\mathbf{x}})]. \tag{22}$$

The right-hand side is positive under Eq. 12. □

A scalar sufficient condition. If an additional moment bound $\mathbb{E}_{\mathbf{x}\sim\mu_h}[V_{\mathbf{x}}] \leq \epsilon^2$ is available, concavity of the logarithm gives

$$H \leq \mathbb{E}_{\mathbf{x}\sim\mu_h}[\log(1+V_{\mathbf{x}})] \leq \log(1+\epsilon^2). \tag{23}$$

Thus $(1-\lambda)G_b + \lambda G_h > \lambda\log(1+\epsilon^2)$ is also sufficient. Since $G_h \geq 0$, the more conservative condition $(1-\lambda)G_b > \lambda\log(1+\epsilon^2)$ is sufficient as well. Neither a small moment bound nor acceptance amplification is assumed by the exact decomposition.

Acceptance and conditional invariance. Eq. 15 shows that steering increases acceptance exactly when $m_{\mathbf{x}} > 1$. By Eq. 16, it preserves the accepted target exactly when $w_{\mathbf{x}} = m_{\mathbf{x}}$ p^+_{ref}-almost surely, equivalently $V_{\mathbf{x}} = 0$. These are distinct properties: acceptance rates alone do not determine the distortion within the accepted set.

A.3 From population targets to the fitted model

For each prompt $\mathbf{x}$, define the population fitting error

$$\delta_\theta(\mathbf{x}) \coloneqq \mathrm{KL}\big(\pi^+_{\mathrm{TS}}(\cdot \mid \mathbf{x}) \,\|\, p_\theta(\cdot \mid \mathbf{x})\big). \tag{24}$$

This measures how closely the fitted model matches the accepted target distribution. It equals the conditional population negative log-likelihood minus the entropy of the target, rather than the empirical training loss. Throughout the pointwise derivations below, we fix $\mathbf{x}$ and suppress conditioning on it when convenient.

Filter acceptance. The fitted model's acceptance probability is

$$\alpha_{p_\theta}(\mathbf{x}) = \sum_{\mathbf{y}\in\mathcal{S}_{\mathbf{x}}} p_\theta(\mathbf{y} \mid \mathbf{x}).$$

Suppose first that $\delta_\theta(\mathbf{x}) < \infty$. Then $p_\theta(\mathbf{y} \mid \mathbf{x}) > 0$ whenever $\pi^+_{\mathrm{TS}}(\mathbf{y} \mid \mathbf{x}) > 0$. Otherwise, a positive-probability term in the KL would be infinite. Since π^+_{TS} assigns total probability one to $\mathcal{S}_{\mathbf{x}}$, it follows that $\alpha_{p_\theta}(\mathbf{x}) > 0$.

We can therefore condition the fitted model on acceptance:

$$p^+_\theta(\mathbf{y} \mid \mathbf{x}) \coloneqq \frac{p_\theta(\mathbf{y} \mid \mathbf{x})\varphi(\mathbf{x},\mathbf{y})}{\alpha_{p_\theta}(\mathbf{x})}.$$

For $\mathbf{y} \in \mathcal{S}_{\mathbf{x}}$, the filter equals one, so

$$p_\theta(\mathbf{y} \mid \mathbf{x}) = \alpha_{p_\theta}(\mathbf{x}) p^+_\theta(\mathbf{y} \mid \mathbf{x}).$$

Substituting this equality into the definition of KL gives

$$\begin{aligned}
\delta_\theta(\mathbf{x}) &= \sum_{\mathbf{y}\in\mathcal{S}_\mathbf{x}} \pi^+_{\mathrm{TS}}(\mathbf{y}\mid\mathbf{x}) \log \frac{\pi^+_{\mathrm{TS}}(\mathbf{y}\mid\mathbf{x})}{p_\theta(\mathbf{y}\mid\mathbf{x})} \\
&= \sum_{\mathbf{y}\in\mathcal{S}_\mathbf{x}} \pi^+_{\mathrm{TS}}(\mathbf{y}\mid\mathbf{x}) \log \frac{\pi^+_{\mathrm{TS}}(\mathbf{y}\mid\mathbf{x})}{\alpha_{p_\theta}(\mathbf{x}) p^+_\theta(\mathbf{y}\mid\mathbf{x})} \\
&= \sum_{\mathbf{y}\in\mathcal{S}_\mathbf{x}} \pi^+_{\mathrm{TS}}(\mathbf{y}\mid\mathbf{x}) \log \frac{\pi^+_{\mathrm{TS}}(\mathbf{y}\mid\mathbf{x})}{p^+_\theta(\mathbf{y}\mid\mathbf{x})} \\
&\qquad - \log \alpha_{p_\theta}(\mathbf{x}) \sum_{\mathbf{y}\in\mathcal{S}_\mathbf{x}} \pi^+_{\mathrm{TS}}(\mathbf{y}\mid\mathbf{x}) \\
&= \mathrm{KL}(\pi^+_{\mathrm{TS}} \| p^+_\theta) - \log \alpha_{p_\theta}(\mathbf{x}). \qquad (25)
\end{aligned}$$

The final equality uses $\sum_{\mathbf{y}\in\mathcal{S}_\mathbf{x}} \pi^+_{\mathrm{TS}}(\mathbf{y}\mid\mathbf{x}) = 1$. Terms with zero target probability are omitted.

Since KL divergence is nonnegative, we obtain the following sequence of implications:

$$\begin{aligned}
\delta_\theta(\mathbf{x}) &\geq -\log \alpha_{p_\theta}(\mathbf{x}), \\
-\delta_\theta(\mathbf{x}) &\leq \log \alpha_{p_\theta}(\mathbf{x}), \\
e^{-\delta_\theta(\mathbf{x})} &\leq \alpha_{p_\theta}(\mathbf{x}), \\
1 - \alpha_{p_\theta}(\mathbf{x}) &\leq 1 - e^{-\delta_\theta(\mathbf{x})}.
\end{aligned}$$

The second line follows by multiplying by -1 and reversing the inequality. The third follows because the exponential function is increasing. The fourth follows by subtracting both sides from one.

Because the filter is binary,

$$\Pr_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}[\varphi(\mathbf{x},\mathbf{y}) = 0] = 1 - \alpha_{p_\theta}(\mathbf{x}).$$

Consequently,

$$\Pr_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}[\varphi(\mathbf{x},\mathbf{y}) = 0] \leq 1 - e^{-\delta_\theta(\mathbf{x})}. \qquad (26)$$

If $\delta_\theta(\mathbf{x}) = \infty$, the right-hand side equals one under the convention $e^{-\infty} = 0$, so the inequality remains valid.

In particular, if $\delta_\theta(\mathbf{x}) < -\log \alpha_{\mathrm{ref}}(\mathbf{x})$, then

$$\alpha_{p_\theta}(\mathbf{x}) \geq e^{-\delta_\theta(\mathbf{x})} > e^{\log \alpha_{\mathrm{ref}}(\mathbf{x})} = \alpha_{\mathrm{ref}}(\mathbf{x}).$$

Thus, this condition guarantees higher filter acceptance than the reference. The argument applies to any target supported on $\mathcal{S}_\mathbf{x}$, including filtered teacher targets.

Benign response preservation. For a benign prompt, $\pi^+_{\mathrm{TS}} = p^+_{\mathrm{ref}}$, and hence

$$\delta_\theta(\mathbf{x}) = \mathrm{KL}(p^+_{\mathrm{ref}} \| p_\theta).$$

We first bound the distance between the fitted model and this target, then account for the change introduced by filtering.

Pinsker's inequality, proved below, states that

$$\mathrm{TV}(P, Q) \leq \sqrt{\frac{\mathrm{KL}(P\|Q)}{2}},$$

where logarithms are natural and $\mathrm{TV}(P,Q) = \frac{1}{2}\sum_\mathbf{y} |P(\mathbf{y}) - Q(\mathbf{y})|$. Applying it with $P = p^+_{\mathrm{ref}}$ and $Q = p_\theta$ gives

$$\begin{aligned}
\mathrm{TV}(p_\theta, p^+_{\mathrm{ref}}) &= \mathrm{TV}(p^+_{\mathrm{ref}}, p_\theta) \\
&\leq \sqrt{\frac{\mathrm{KL}(p^+_{\mathrm{ref}}\|p_\theta)}{2}} \\
&= \sqrt{\frac{\delta_\theta(\mathbf{x})}{2}}.
\end{aligned}$$

Only the arguments of total variation are exchanged here. The direction of the KL remains $p^+_{\text{ref}} \| p_\theta$.

Next, expand the total variation distance to the reference:

$$\begin{aligned}
\text{TV}(p_\theta, p_{\text{ref}}) &= \frac{1}{2} \sum_{\mathbf{y} \in \mathcal{Y}} |p_\theta(\mathbf{y} \mid \mathbf{x}) - p_{\text{ref}}(\mathbf{y} \mid \mathbf{x})| \\
&= \frac{1}{2} \sum_{\mathbf{y} \in \mathcal{Y}} \Big| \big(p_\theta(\mathbf{y} \mid \mathbf{x}) - p^+_{\text{ref}}(\mathbf{y} \mid \mathbf{x})\big) + \big(p^+_{\text{ref}}(\mathbf{y} \mid \mathbf{x}) - p_{\text{ref}}(\mathbf{y} \mid \mathbf{x})\big) \Big| \\
&\leq \frac{1}{2} \sum_{\mathbf{y} \in \mathcal{Y}} \big|p_\theta(\mathbf{y} \mid \mathbf{x}) - p^+_{\text{ref}}(\mathbf{y} \mid \mathbf{x})\big| + \frac{1}{2} \sum_{\mathbf{y} \in \mathcal{Y}} \big|p^+_{\text{ref}}(\mathbf{y} \mid \mathbf{x}) - p_{\text{ref}}(\mathbf{y} \mid \mathbf{x})\big| \\
&= \text{TV}(p_\theta, p^+_{\text{ref}}) + \text{TV}(p^+_{\text{ref}}, p_{\text{ref}}).
\end{aligned}$$

The inequality applies $|a + b| \leq |a| + |b|$ separately to each response.

Combining this calculation with Pinsker's inequality and $\text{TV}(p^+_{\text{ref}}, p_{\text{ref}}) = 1 - \alpha_{\text{ref}}(\mathbf{x})$ from Eq. 14 yields

$$\text{TV}(p_\theta, p_{\text{ref}}) \leq \sqrt{\frac{\delta_\theta(\mathbf{x})}{2}} + 1 - \alpha_{\text{ref}}(\mathbf{x}). \tag{27}$$

The first term bounds the error in fitting the accepted target. The second is the exact change caused by filtering the native reference distribution.

From total variation to response scores. We next justify why total variation bounds changes in any score taking values in $[0, 1]$.

For arbitrary distributions P, Q on $\mathcal{Y}$, define

$$\mathcal{A} \coloneqq \{\mathbf{y} \in \mathcal{Y} : P(\mathbf{y}) \geq Q(\mathbf{y})\}.$$

Since both distributions sum to one,

$$\sum_{\mathbf{y} \in \mathcal{Y}} (P(\mathbf{y}) - Q(\mathbf{y})) = 0.$$

Therefore, the total positive difference equals the magnitude of the total negative difference:

$$\begin{aligned}
\sum_{\mathbf{y} \in \mathcal{A}} (P(\mathbf{y}) - Q(\mathbf{y})) &= - \sum_{\mathbf{y} \notin \mathcal{A}} (P(\mathbf{y}) - Q(\mathbf{y})) \\
&= \text{TV}(P, Q).
\end{aligned}$$

Fix $\mathbf{x}$ and let $s(\mathbf{x}, \mathbf{y}) \in [0, 1]$. On $\mathcal{A}$, the difference $P - Q$ is nonnegative, so multiplying it by s cannot increase it. On the complement, the difference is negative, so its contribution to an upper bound can be discarded. Hence,

$$\begin{aligned}
\mathbb{E}_{\mathbf{y} \sim P}[s(\mathbf{x}, \mathbf{y})] - \mathbb{E}_{\mathbf{y} \sim Q}[s(\mathbf{x}, \mathbf{y})] &= \sum_{\mathbf{y} \in \mathcal{Y}} s(\mathbf{x}, \mathbf{y})(P(\mathbf{y}) - Q(\mathbf{y})) \\
&\leq \sum_{\mathbf{y} \in \mathcal{A}} (P(\mathbf{y}) - Q(\mathbf{y})) \\
&= \text{TV}(P, Q).
\end{aligned}$$

Conversely, the contribution on $\mathcal{A}$ is nonnegative, and multiplying a negative difference by $s \leq 1$ makes it no smaller. Thus,

$$\begin{aligned}
\mathbb{E}_{\mathbf{y} \sim P}[s(\mathbf{x}, \mathbf{y})] - \mathbb{E}_{\mathbf{y} \sim Q}[s(\mathbf{x}, \mathbf{y})] &\geq \sum_{\mathbf{y} \notin \mathcal{A}} (P(\mathbf{y}) - Q(\mathbf{y})) \\
&= - \text{TV}(P, Q).
\end{aligned}$$

Combining the upper and lower bounds gives

$$|\mathbb{E}_{\mathbf{y} \sim P}[s(\mathbf{x}, \mathbf{y})] - \mathbb{E}_{\mathbf{y} \sim Q}[s(\mathbf{x}, \mathbf{y})]| \leq \text{TV}(P, Q).$$

Applying this result to $P = p_\theta(\cdot \mid \mathbf{x})$ and $Q = p_{\text{ref}}(\cdot \mid \mathbf{x})$, followed by Eq. 27, gives

$$\begin{aligned} &|\mathbb{E}_{\mathbf{y}\sim p_\theta}[s(\mathbf{x},\mathbf{y})] - \mathbb{E}_{\mathbf{y}\sim p_{\text{ref}}}[s(\mathbf{x},\mathbf{y})]| \\ &\qquad \le 1 - \alpha_{\text{ref}}(\mathbf{x}) + \sqrt{\frac{\delta_\theta(\mathbf{x})}{2}}. \end{aligned} \tag{28}$$

A correctness indicator is one possible choice of s.

Averaging over benign prompts. For this calculation, define the per-prompt score difference

$$d_\theta(\mathbf{x}) := \mathbb{E}_{\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}[s(\mathbf{x},\mathbf{y})] - \mathbb{E}_{\mathbf{y}\sim p_{\text{ref}}(\cdot|\mathbf{x})}[s(\mathbf{x},\mathbf{y})].$$

Writing out the joint expectations shows that their difference equals the average of these per-prompt differences:

$$\begin{aligned} &\mathbb{E}_{\substack{\mathbf{x}\sim\mu_b \\ \mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}}[s(\mathbf{x},\mathbf{y})] - \mathbb{E}_{\substack{\mathbf{x}\sim\mu_b \\ \mathbf{y}\sim p_{\text{ref}}(\cdot|\mathbf{x})}}[s(\mathbf{x},\mathbf{y})] \\ &\quad = \sum_{\mathbf{x}} \mu_b(\mathbf{x}) \sum_{\mathbf{y}} s(\mathbf{x},\mathbf{y}) \big(p_\theta(\mathbf{y} \mid \mathbf{x}) - p_{\text{ref}}(\mathbf{y} \mid \mathbf{x})\big) \\ &\quad = \sum_{\mathbf{x}} \mu_b(\mathbf{x}) d_\theta(\mathbf{x}). \end{aligned}$$

Here, sums over $\mathbf{x}$ run over the finite support of μ_b.

Applying the triangle inequality to the finite sum and then using Eq. 28 gives

$$\begin{aligned} \left|\sum_{\mathbf{x}} \mu_b(\mathbf{x}) d_\theta(\mathbf{x})\right| &\le \sum_{\mathbf{x}} \mu_b(\mathbf{x}) |d_\theta(\mathbf{x})| \\ &\le \sum_{\mathbf{x}} \mu_b(\mathbf{x}) \left(1 - \alpha_{\text{ref}}(\mathbf{x}) + \sqrt{\frac{\delta_\theta(\mathbf{x})}{2}}\right) \\ &= \mathbb{E}_{\mathbf{x}\sim\mu_b}[1 - \alpha_{\text{ref}}(\mathbf{x})] + \mathbb{E}_{\mathbf{x}\sim\mu_b}\left[\sqrt{\frac{\delta_\theta(\mathbf{x})}{2}}\right]. \end{aligned}$$

It remains to bound the average square root. For finite average fitting error, the Cauchy–Schwarz inequality gives

$$\begin{aligned} \left(\sum_{\mathbf{x}} \mu_b(\mathbf{x}) \sqrt{\frac{\delta_\theta(\mathbf{x})}{2}}\right)^2 &= \left(\sum_{\mathbf{x}} \sqrt{\mu_b(\mathbf{x})} \sqrt{\frac{\mu_b(\mathbf{x})\delta_\theta(\mathbf{x})}{2}}\right)^2 \\ &\le \left(\sum_{\mathbf{x}} \mu_b(\mathbf{x})\right) \left(\sum_{\mathbf{x}} \frac{\mu_b(\mathbf{x})\delta_\theta(\mathbf{x})}{2}\right) \\ &= \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\mu_b}[\delta_\theta(\mathbf{x})]. \end{aligned}$$

The last equality uses $\sum_{\mathbf{x}} \mu_b(\mathbf{x}) = 1$. Taking square roots gives

$$\mathbb{E}_{\mathbf{x}\sim\mu_b}\left[\sqrt{\frac{\delta_\theta(\mathbf{x})}{2}}\right] \le \sqrt{\frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\mu_b}[\delta_\theta(\mathbf{x})]}.$$

This is also Jensen's inequality for the concave square-root function. If the average fitting error is infinite, the inequality holds trivially with an infinite right-hand side.

Substituting this bound into the preceding calculation yields

$$\begin{aligned} &\left|\mathbb{E}_{\substack{\mathbf{x}\sim\mu_b \\ \mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}}[s(\mathbf{x},\mathbf{y})] - \mathbb{E}_{\substack{\mathbf{x}\sim\mu_b \\ \mathbf{y}\sim p_{\text{ref}}(\cdot|\mathbf{x})}}[s(\mathbf{x},\mathbf{y})]\right| \\ &\qquad \le \mathbb{E}_{\mathbf{x}\sim\mu_b}[1 - \alpha_{\text{ref}}(\mathbf{x})] + \sqrt{\frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\mu_b}[\delta_\theta(\mathbf{x})]}. \end{aligned} \tag{29}$$

Derivation of Pinsker's inequality. For completeness, we derive the KL-to-TV inequality used above. Let P, Q be distributions on a finite space and define

$$\mathcal{A} \coloneqq \{\mathbf{y} : P(\mathbf{y}) \geq Q(\mathbf{y})\}, \qquad a \coloneqq P(\mathcal{A}), \qquad b \coloneqq Q(\mathcal{A}).$$

The positive-difference identity established above gives

$$a - b = \sum_{\mathbf{y} \in \mathcal{A}} (P(\mathbf{y}) - Q(\mathbf{y})) = \mathrm{TV}(P, Q),$$

so $a \geq b$.

First suppose $0 < a, b < 1$. Conditioning on $\mathcal{A}$ and its complement decomposes the KL as

$$\begin{aligned} \mathrm{KL}(P\|Q) &= a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b} \\ &\quad + a\, \mathrm{KL}(P(\cdot \mid \mathcal{A}) \,\|\, Q(\cdot \mid \mathcal{A})) \\ &\quad + (1-a)\, \mathrm{KL}(P(\cdot \mid \mathcal{A}^c) \,\|\, Q(\cdot \mid \mathcal{A}^c)) \\ &\geq a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b}. \end{aligned}$$

The decomposition follows by writing $P(\mathbf{y}) = aP(\mathbf{y} \mid \mathcal{A})$ and $Q(\mathbf{y}) = bQ(\mathbf{y} \mid \mathcal{A})$ on $\mathcal{A}$, and using the analogous factorization on $\mathcal{A}^c$. The inequality discards two nonnegative conditional KL terms.

For fixed $b \in (0, 1)$, define

$$F_b(u) \coloneqq u \log \frac{u}{b} + (1-u) \log \frac{1-u}{1-b}.$$

Direct differentiation gives

$$F_b(b) = 0, \qquad F_b'(b) = 0, \qquad F_b''(u) = \frac{1}{u} + \frac{1}{1-u} = \frac{1}{u(1-u)} \geq 4,$$

where the last inequality uses $u(1-u) \leq 1/4$. For $u \geq b$, integrating this bound gives

$$F_b'(u) = F_b'(b) + \int_b^u F_b''(v)\, dv \geq 4(u - b).$$

Integrating once more, using $a \geq b$, yields

$$F_b(a) = F_b(b) + \int_b^a F_b'(u)\, du \geq \int_b^a 4(u-b)\, du = 2(a-b)^2.$$

Consequently,

$$\mathrm{KL}(P\|Q) \geq F_b(a) \geq 2(a-b)^2 = 2\, \mathrm{TV}(P, Q)^2.$$

Rearranging establishes

$$\mathrm{TV}(P, Q) \leq \sqrt{\frac{\mathrm{KL}(P\|Q)}{2}}.$$

The case $a = 1$ with $b \in (0, 1)$ follows by continuity. If $b = 0$ and $a > 0$, then $\mathrm{KL}(P\|Q) = \infty$, and the inequality is trivial. If $a = b = 0$ or $b = 1$, then $a - b = 0$ and the TV is zero. These cases cover the remaining boundaries.

Interpretation. High benign acceptance and small average population fitting error suffice to preserve average benign response scores. The inequalities remain valid when fitting error is large, but they may then be uninformative. This analysis does not establish that finite-data fine-tuning achieves small population fitting error or that preservation transfers to a different prompt distribution. The acceptance guarantee is measured by the fixed filter φ.

B THINKSAFE

B.1 Sampling details

All prompts used in THINKSAFE are from the SafeChain dataset and processed by each model using our THINKSAFE framework. For Qwen3 model family, we sample one response for each prompt with top-p 0.95, top-k 20, temperature 0.6, and maximum token limit 16,384. For DeepSeek-R1-Distill family, we use the same hyperparameter except that top-k is set to 0. Responses that are classified as unsafe by the Llama-Guard-3, denoted as φ, are excluded from the analysis. Table 3 shows the ratio of filtered samples per model. Larger Qwen3 models (4B and 8B) retain over 99% of both benign and harmful samples, while the smaller Qwen3-0.6B and Qwen3-1.7B filter out 4.84% and 3.29% of harmful samples, respectively. The R1-Distill-1.5B model exhibits substantially higher filtering rates. Overall, larger models tend to preserve a greater portion of the original data, suggesting more stable and consistent response distributions after filtering.

Table 3: Filtered ratio (%) per model.

Sample category	Qwen3				DeepSeek-R1-Distill		
	0.6B	1.7B	4B	8B	1.5B	7B	8B
Benign	1.07	0.93	0.63	0.74	11.01	2.69	0.81
Harmful	4.84	3.29	0.35	0.16	12.86	2.93	0.48

B.2 Statistics

We report the statistics of the responses generated by THINKSAFE in Figs. 9 and 10. Here, N_h and N_b denote the numbers of harmful and benign prompts, respectively, while μ_h and μ_b represent the average response lengths (in tokens) for harmful and benign queries.

Across both the Qwen3 and R1-distilled model series, benign responses consistently exhibit longer generation lengths than harmful ones, reflecting the presence of more detailed reasoning traces. Moreover, as model size increases, both harmful and benign responses tend to become longer and more stable in distribution.

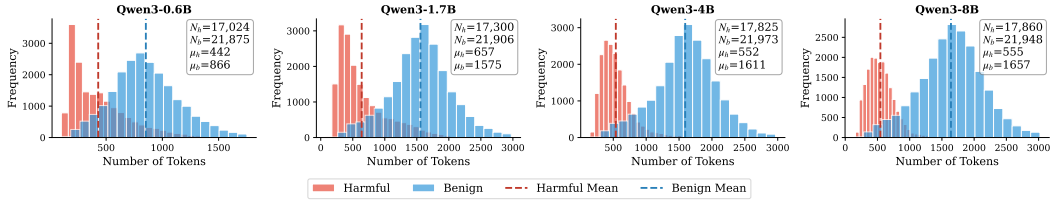


Figure 9: Statistics of THINKSAFE in Qwen3 model series. Top 1% outliers by length are excluded for better interpretability.

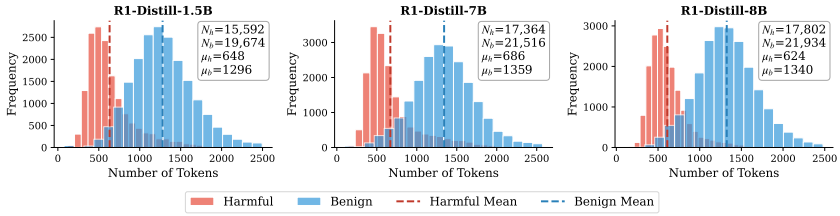


Figure 10: Statistics of THINKSAFE in DeepSeek-R1-Distill model series. Top 1% outliers by length are excluded for better interpretability.

C Experimental Details

For the AIME 2024, GSM8K, MATH500, and GPQA datasets, we use the SkyThought [34] library to evaluate models. We sample 8 responses for each prompt using dataset-specific prompts and report the average pass@1. For the Qwen model family, we use a temperature of 0.6, top-p of 0.95, and

top-k of 20, with a maximum token limit of 32,768. For the DeepSeek-R1-Distill family, we use a temperature of 0.6 and top-p of 0.95, with a maximum token limit of 32,768.

AIME 2024

Please reason step by step, and put your final answer within \\boxed{{}}. `{problem}`

GSM8K

Given the following problem, reason and give a final answer to the problem.
Problem: `{question}`
Your response should end with "The final answer is [answer]" where [answer] is the response to the problem.

GPQA

Return your final response within \\boxed{{}} and only include the letter choice (A, B, C, or D) as your final response. `{problem}`

MATH500

Please reason step by step, and put your final answer within \\boxed{{}}. `{problem}`

D Baseline Details

Table 4: Detailed hyperparameters for baselines.

Method	Epochs	Source	Sample size
DirectRefusal	5	WildJailbreak	1,000
SafeChain	3	WildJailbreak	40,000
STAR-1	5	Mixture of harmful datasets (See 36 for details)	1,000
SafeKey	5	STAR-1	1,000
SafePath	4	WildJailbreak	400
THINKSAFE	3	SafeChain	Varies by scale

To ensure consistency and reproducibility, we adopt the same hyperparameters as specified in the original papers for all baselines, as summarized in Table 4. For THINKSAFE, since the sample size varies across model scales, we refer readers to Figs. 9 and 10 for detailed configurations.

E Additional Experiments

In this section, we present additional experiments to empirically support the effectiveness of the proposed THINKSAFE.

E.1 Necessity of safety reasoning on Qwen families

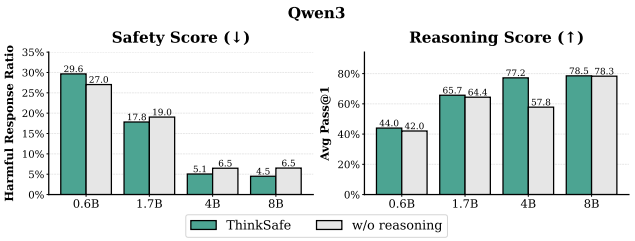


Figure 11: Ablation of safety reasoning in Qwen3 model series.

To further analyze the effect of reasoning on the Qwen family, we ablate reasoning traces from refusal responses $\mathbf{y}_h^{(i)}$, while retaining full chain-of-thought for benign responses $\mathbf{y}_b^{(i)}$. Both types of responses are generated by each model in the Qwen series. As shown in Fig. 11, removing reasoning traces consistently degrades both safety and reasoning performance. In particular, the Qwen3-4B model exhibits a substantial drop in reasoning accuracy, with Avg Pass@1 decreasing from 77.2% to 57.8%, accompanied by an increase in harmful response ratio. These results indicate that explicit reasoning plays a critical role not only in maintaining reasoning capability but also in supporting safety-aligned behavior.

E.2 Cross-model distillation for larger models

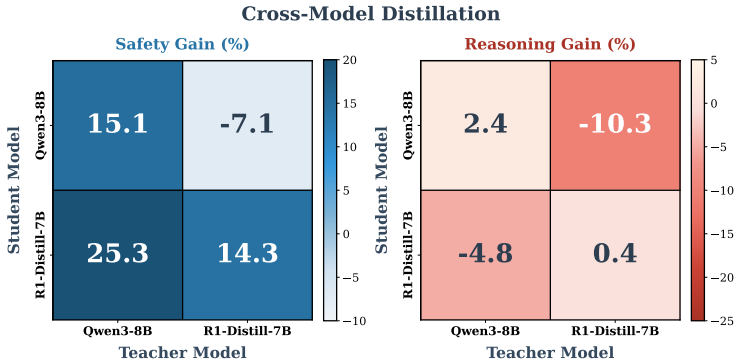


Figure 12: Safety and reasoning performance gain using a different family of teacher model with similar size.

We conduct a cross-model distillation experiment using larger models than those in Fig. 6, namely Qwen3-8B and R1-Distill-Qwen-7B. Fig. 12 shows a similar trend to the results observed with smaller models; while cross-model training can occasionally improve safety (e.g., DeepSeek-R1-Distill achieves a 25.3% safety gain when trained on Qwen3-8B data), it consistently leads to notable degradation in reasoning performance across all reciprocal training settings.

E.3 Refusal steering strategies

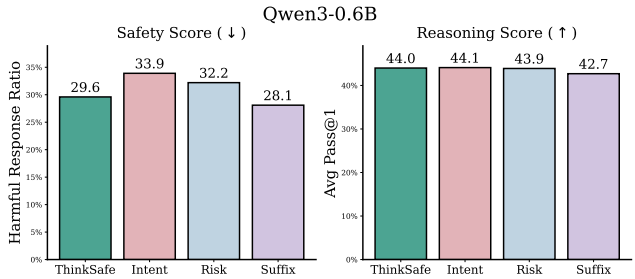


Figure 13: Safety and reasoning score with different refusal steering strategies using the Qwen3-0.6B model.

Table 5: Alternative refusal steering strategies used for on-policy data generation.

Strategy	Prompt Template
THINKSAFE	The following prompt is harmful. You should refuse to answer the prompt.\n\n Prompt:{prompt}
Suffix	{prompt}\n\n The above prompt is harmful. You should refuse to answer the prompt.
Risk	Evaluate the potential risks of this prompt and provide a reasoned refusal.\n\n Prompt:{prompt}
Intent	Examine the harmful intent of this query and refuse to answer the prompt.\n\n Prompt:{prompt}

We further investigate the robustness of THINKSAFE on the Qwen3-0.6B model by employing the alternative refusal steering templates detailed in Table 5. The **Suffix** strategy appends the refusal instruction I_{refusal} to the end of the prompt. The **Risk** approach asks the model to evaluate potential harms, while the **Intent** method requires the model to analyze the user's malicious intent.

As shown in Fig. 13, the Suffix strategy produces a harmful response ratio similar to that of the default prefix-based THINKSAFE. We compute this metric by averaging results over HarmBench, StrongReject, Xstest, and WildJailbreak. This similarity indicates that where the refusal instruction appears is not especially important, as straightforward refusal instructions work well regardless of placement. In contrast, the Risk and Intent strategies lead to noticeably worse safety outcomes, with higher harmful response ratios. We attribute this gap to the added complexity of these instructions. By asking the model to carry out extra reasoning steps instead of issuing a direct refusal, these prompts may weaken the strength of the safety constraint. Importantly, overall reasoning performance remains stable across all four strategies, as measured by average pass@1 on AIME2024, GSM8K, MATH500, and GPQA. This stability supports our central claim. Because all variants depend on self-generated outputs from the model itself rather than learning from an external model, the model's core reasoning ability remains intact even when the safety approach changes.

E.4 Filtering with WildGuard

To assess whether the effectiveness of THINKSAFE depends on the specific characteristics of the filtering model, we conduct an ablation study using WildGuard instead of Llama-Guard-3. We denote this variant as THINKSAFE + WG, which employs WildGuard to filter the self-generated safety data. The results in Table 6 and Table 7 demonstrate remarkable stability in performance and confirm that our framework is robust to the choice of safety classifier. For example, the WildGuard variant of the Qwen3-4B model (THINKSAFE + WG) achieves safety scores nearly identical to the baseline while preserving superior reasoning capabilities. This consistency reinforces our core hypothesis that the success of THINKSAFE arises from the refusal steering mechanism itself rather than from overfitting to a specific reward model. By successfully filtering self-generated traces with a completely different guard model, we demonstrate that the elicited safety behaviors are generalized and transferable.

E.5 Experiments with Larger Models

To verify that THINKSAFE continues to scale beyond the 8B regime reported in Table 1, we additionally apply our framework to Qwen3-14B and Qwen3-32B, with results summarized in Table 8. THINKSAFE consistently achieves the most favorable safety-reasoning trade-off at both scales. On Qwen3-14B, it reduces average harmfulness from 17.44 to **3.04** (a $\sim$5.7$\times$ reduction) while preserving reasoning within 0.44 points of the initial model (82.56 $\rightarrow$ 82.12). On Qwen3-32B, THINKSAFE cuts average harmfulness from 19.72 to **5.58** and simultaneously *improves* average reasoning from 82.69 to **82.85**, outperforming all baselines on both axes. In contrast, teacher-distillation with SafeChain degrades reasoning substantially (79.06 at 14B and 79.82 at 32B), and DirectRefusal exhibits severe over-refusal on benign XSTest prompts (35.20% and 39.20%, respectively). These results confirm that the KL-optimal target p^{+}_{ref} established in Sec. 4 remains empirically realizable at larger scales, and that the advantages of self-generation over external supervision *grow* rather than diminish as student capacity increases.

Table 6: Results on **Qwen3** models. Safety reports harmful-response rates on HarmBench, StrongReject, and WildJailbreak, and over-refusal on benign XSTest prompts (↓); their unweighted mean is the safety average. Reasoning reports average pass@1 from eight responses per prompt (↑); its average is the unweighted mean of four benchmarks. All scores are percentages; averages are computed from the displayed component scores. Best and second-best distinct values within each model size are **bold** and underlined; ties share a rank.

		Safety (↓)					Reasoning (Avg pass@1, ↑)				
		Harmfulness			Over-refusal						
Size	Method	Harm Bench	Strong Reject	Wild Jailbreak	XSTest	Safety Average	AIME 2024	GSM8K	MATH 500	GPQA	Reasoning Average
0.6B	Initial	68.44	66.45	52.80	5.20	48.22	**10.42**	72.51	71.73	25.13	**44.95**
	DirectRefusal	43.85	**11.82**	**36.30**	83.60	43.89	5.83	64.30	67.53	24.81	40.62
	SafeChain	58.64	72.84	49.60	**0.00**	45.27	4.58	68.68	62.42	23.74	39.86
	STAR-1	56.64	38.02	50.60	22.40	41.92	6.25	68.15	68.17	24.18	41.69
	SafePath	67.61	60.06	52.80	4.40	46.22	7.92	71.26	71.77	**26.07**	44.26
	SafeKey	60.96	48.88	52.75	18.40	45.25	5.42	71.58	66.17	24.94	42.03
	THINKSAFE	40.37	33.87	37.95	6.40	**29.65**	9.58	72.36	70.65	23.30	43.97
	THINKSAFE+WG	**39.04**	35.46	37.55	7.20	29.81	9.58	**72.82**	**72.00**	24.81	44.80
1.7B	Initial	52.66	36.10	51.10	**1.20**	35.27	44.58	84.31	88.85	41.73	64.87
	DirectRefusal	38.54	**5.75**	35.75	61.60	35.41	43.75	82.78	88.10	41.29	63.98
	SafeChain	47.34	57.51	43.85	1.60	37.58	34.58	**85.29**	85.72	38.13	60.93
	STAR-1	37.38	7.67	46.60	10.80	25.61	**46.25**	84.38	88.30	41.16	**65.02**
	SafePath	54.15	36.42	49.30	**1.20**	35.27	43.33	84.33	88.32	**42.42**	64.60
	SafeKey	46.84	18.21	48.85	8.80	30.68	38.33	84.31	88.12	40.03	62.70
	THINKSAFE	**28.74**	9.58	29.20	2.00	17.38	44.17	83.80	**89.05**	40.53	64.39
	THINKSAFE+WG	28.90	7.99	**29.00**	2.00	**16.97**	45.00	83.43	88.98	41.04	64.61
4B	Initial	38.21	8.31	43.00	**0.80**	22.58	67.50	84.69	93.43	52.27	74.47
	DirectRefusal	33.06	3.19	36.20	32.00	26.11	68.33	82.58	93.20	53.03	74.29
	SafeChain	43.69	41.21	39.65	2.00	31.64	62.08	89.59	93.03	51.01	73.93
	STAR-1	33.72	5.75	35.15	6.80	20.36	62.50	90.97	93.05	51.96	74.62
	SafePath	37.71	7.35	42.45	1.60	22.28	72.08	84.45	93.33	53.54	75.85
	SafeKey	32.39	3.19	32.95	**0.80**	17.33	67.08	**91.79**	92.87	51.83	75.89
	THINKSAFE	9.63	**0.32**	7.45	2.80	5.05	73.33	88.06	93.53	**53.79**	77.18
	THINKSAFE+WG	**9.47**	**0.32**	**7.25**	2.40	**4.86**	**75.42**	88.15	**93.55**	53.60	**77.68**
8B	Initial	35.05	4.47	38.35	**0.40**	19.57	74.17	85.28	**94.18**	50.69	76.08
	DirectRefusal	24.42	1.92	28.05	37.60	23.00	**74.58**	84.31	93.63	59.41	77.98
	SafeChain	41.20	38.95	36.42	1.20	29.44	70.00	**92.98**	93.53	58.21	78.68
	STAR-1	24.42	1.28	29.25	6.80	15.44	72.50	90.29	93.73	57.83	78.59
	SafePath	35.22	6.71	39.45	1.20	20.65	**74.58**	84.89	93.85	**61.24**	78.64
	SafeKey	26.91	4.79	28.80	8.80	17.33	70.00	92.44	93.30	59.91	**78.91**
	THINKSAFE	**9.14**	**0.32**	7.35	1.20	**4.50**	72.92	88.00	93.10	59.67	78.42
	THINKSAFE+WG	9.63	**0.32**	**7.05**	3.20	5.05	73.33	87.96	93.65	60.29	78.81

E.6 Controlled Comparison with SafePath

Throughout the experiment and analyses, we strictly follow the original training configuration of each baseline to ensure a fair comparison under its intended setting. However, since the number of training examples differs across baselines (see Table 4), we further conduct a controlled comparison with matched data scale. We use SafePath as the reference, as it uses the smallest training set among the baselines, and train it with the same prompt set as THINKSAFE. Following the original SafePath protocol, we add the safety primer "<think> Let's think about safety first." only to harmful prompts.

As shown in Table 9, the controlled SafePath baseline remains substantially less effective under the matched data scale. THINKSAFE preserves reasoning while more than halving the average harm rate (4.50% vs 11.50%) compared to the controlled SafePath, confirming that native self-generated refusal traces provide superior safety grounding.

F Online Learning Baseline Details

F.1 GRPO Objective

The objective function for GRPO is formulated to optimize the policy without a separate value function, instead estimating advantages in a group-relative manner as introduced in Shao et al. [29]. Given a dataset $\mathcal{D}$ of prompts $\mathbf{x}$, GRPO samples a group of G candidate responses $\{\mathbf{y}_i\}_{i=1}^{G}$ from a behavior policy $\pi_{\theta_{\text{old}}}$, while constraining updates to remain close to a fixed reference policy π_{ref}. Then, the advantage $\hat{A}_i$ for each response $\mathbf{y}_i$ in a group of size G is computed by normalizing the

Table 7: Results on **DeepSeek-R1-Distill** models. Safety reports harmful-response rates on HarmBench, StrongReject, and WildJailbreak, and over-refusal on benign XSTest prompts (↓); their unweighted mean is the safety average. Reasoning reports average pass@1 from eight responses per prompt (↑); its average is the unweighted mean of four benchmarks. All scores are percentages; averages are computed from the displayed component scores. Best and second-best distinct values within each model size are **bold** and underlined; ties share a rank.

		Safety (↓)					Reasoning (Avg pass@1, ↑)				
		Harmfulness			Over-refusal						
Size	Method	Harm Bench	Strong Reject	Wild Jailbreak	XSTest	Safety Average	AIME 2024	GSM8K	MATH 500	GPQA	Reasoning Average
1.5B	Initial	67.28	82.11	51.55	**0.00**	50.24	21.25	82.42	79.45	31.94	53.77
	DirectRefusal	66.11	82.75	50.45	8.40	51.93	19.17	81.06	78.55	32.70	52.87
	SafeChain	59.30	76.68	46.95	0.40	45.83	24.17	80.47	81.25	28.28	53.54
	STAR-1	62.79	77.00	49.65	1.20	47.66	17.08	81.38	79.07	31.25	52.20
	SafePath	65.28	82.43	51.80	0.40	49.98	24.17	82.37	79.57	32.51	54.66
	SafeKey	58.80	**73.16**	47.65	3.60	45.80	19.58	81.25	78.02	28.72	51.89
	THINKSAFE	**52.99**	74.12	40.50	1.20	**42.20**	**32.92**	**82.58**	82.50	31.19	**57.30**
	THINKSAFE+WG	54.82	76.68	**39.95**	1.20	43.16	30.00	82.36	**82.55**	**32.89**	56.95
7B	Initial	56.98	63.58	53.15	1.20	43.73	49.58	90.32	90.18	46.65	69.18
	DirectRefusal	52.33	**33.55**	50.20	43.60	44.92	47.50	88.27	89.82	44.95	67.64
	SafeChain	51.00	54.63	45.85	0.40	37.97	49.17	89.75	91.50	46.78	69.30
	STAR-1	52.99	47.92	48.75	2.40	38.02	45.83	90.32	90.58	46.02	68.19
	SafePath	55.15	64.86	52.65	**0.00**	43.17	52.08	89.71	90.62	46.15	69.64
	SafeKey	45.35	33.87	45.75	7.20	33.04	43.75	90.58	89.90	**47.29**	67.88
	THINKSAFE	40.20	41.85	35.40	0.40	29.46	51.25	90.10	91.90	45.20	69.61
	THINKSAFE+WG	**40.03**	38.98	**34.25**	0.80	**28.52**	**52.50**	**90.98**	**92.05**	46.97	**70.63**
8B	Initial	52.33	53.99	49.70	**0.40**	39.11	47.50	**87.74**	87.38	**48.11**	**67.68**
	DirectRefusal	32.39	**0.64**	32.60	50.00	28.91	40.00	83.26	85.00	43.50	62.94
	SafeChain	44.52	46.33	42.45	1.60	33.73	41.67	86.06	86.50	42.05	64.07
	STAR-1	**21.26**	3.51	**17.60**	12.00	**13.59**	40.42	87.28	86.65	43.69	64.51
	SafePath	47.51	51.44	50.20	**0.40**	37.39	43.33	87.45	87.43	47.41	66.41
	SafeKey	32.72	11.82	30.85	8.00	20.85	35.83	87.41	85.80	42.49	62.88
	THINKSAFE	27.08	26.52	21.15	1.60	19.09	**48.75**	87.55	**87.70**	46.28	67.57
	THINKSAFE+WG	27.24	26.20	22.00	1.60	19.26	45.42	87.62	86.82	45.14	66.25

Table 8: Results on **Qwen3-14B and Qwen3-32B** models. We evaluate safety across three benchmarks (HarmBench, StrongReject, WildJailbreak) by reporting the ratio of harmful responses (↓). Over-refusal is measured by the refusal rate (↓) on benign XSTest prompts. For reasoning tasks, we sample 8 trajectories per prompt and report the average pass@1 (↑). Best results are **bolded**; second best are underlined.

		Safety (↓)					Reasoning (Avg pass@1, ↑)				
		Harmfulness			Over-refusal						
Size	Method	Harm Bench	Strong Reject	Wild Jailbreak	XSTest	Safety Average	AIME 2024	GSM8k	MATH 500	GPQA	Reasoning Average
14B	Initial	29.90	3.19	35.05	1.60	17.44	**79.17**	92.58	**94.67**	**63.83**	**82.56**
	DirectRefusal	21.59	0.64	23.15	35.20	20.15	77.08	92.69	94.47	62.82	81.77
	SafeChain	46.01	46.65	38.50	**1.20**	33.09	69.17	91.55	94.27	61.24	79.06
	THINKSAFE	**6.31**	**0.00**	**3.45**	2.40	**3.04**	77.50	**93.20**	94.27	63.51	82.12
32B	Initial	35.71	5.11	36.45	1.60	19.72	79.17	87.82	**95.15**	**68.62**	82.69
	DirectRefusal	18.44	**0.32**	21.05	39.20	19.75	**79.58**	87.51	95.03	67.36	82.37
	SafeChain	42.86	46.65	36.45	**0.80**	31.69	72.92	**91.79**	94.35	60.23	79.82
	THINKSAFE	**10.47**	**0.32**	**7.15**	4.40	**5.58**	**79.58**	89.13	94.95	67.74	**82.85**

rewards r_i against the group's mean and standard deviation:

$$\hat{A}_i = \frac{r_i - \text{mean}(\{r_1, \ldots, r_G\})}{\text{std}(\{r_1, \ldots, r_G\})} \tag{30}$$

These advantages are used in a clipped objective with importance sampling, together with a KL penalty term:

$$\mathcal{J}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(\mathbf{y}_i \mid \mathbf{x})}{\pi_{\theta_{\text{old}}}(\mathbf{y}_i \mid \mathbf{x})}\hat{A}_i, \text{clip}\left(\frac{\pi_\theta(\mathbf{y}_i \mid \mathbf{x})}{\pi_{\theta_{\text{old}}}(\mathbf{y}_i \mid \mathbf{x})}, 1-\epsilon, 1+\epsilon\right)\hat{A}_i\right) - \beta\,\text{KL}\left(\pi_\theta \,\|\, \pi_{\text{ref}}\right)\right)\right]$$

Table 9: Controlled comparison with SafePath using Qwen3-8B.

Method	Safety (↓)					Reasoning (Avg pass@1, ↑)				
	Harmfulness			Over-refusal	Safety Average					
	Harm Bench	Strong Reject	Wild Jailbreak	XSTest		AIME 2024	GSM8k	MATH 500	GPQA	Reasoning Average
SafePath	21.76	2.24	20.00	2.00	11.50	**73.33**	85.57	92.93	**61.74**	78.39
THINKSAFE	**9.14**	**0.32**	**7.35**	**1.20**	**4.50**	72.92	**88.00**	**93.10**	59.97	**78.50**

F.2 Safety Reward

The safety reward r_{safety} is derived from the output logits of the safety guard model. Specifically, given a prompt $\mathbf{x}$ and a sampled response $\mathbf{y}$, we extract the log probabilities of the tokens corresponding to "`safe`" ($c = 1$) and "`unsafe`" ($c = 0$) from the guard model's prediction. The final reward is computed by applying a softmax function over these two token probabilities to represent the likelihood of the response being safe:

$$r_{\text{safety}}(\mathbf{x}, \mathbf{y}) = \frac{p_\varphi(c = 1 \mid \mathbf{x}, \mathbf{y})}{p_\varphi(c = 1 \mid \mathbf{x}, \mathbf{y}) + p_\varphi(c = 0 \mid \mathbf{x}, \mathbf{y})} \tag{31}$$

The r_{safety} ensures a continuous reward signal between 0 and 1, reflecting the guard model's confidence in the safety of the generated trace. Subsequently, the r_{safety} is combined with the r_{format} to constitute the total reward r, which is then utilized in Eq. 30.

F.3 Format Reward

The format reward, r_{format}, strictly enforces the structural integrity of the reasoning traces. For standard models, we assign a reward of 1 if the response contains exactly one pair of `<think>` and `</think>` tags in the correct order, and 0 otherwise. For the DeepSeek-R1-Distill family, which omits the opening tag by default, we adapt this criterion to require exactly one occurrence of the closing `</think>` tag and no opening `<think>` tag. This ensures that the model maintains a consistent chain-of-thought structure during the online RL process.

F.4 OPSD Objective

OPSD [42] trains a student to minimize the token-level reverse KL divergence against a teacher distribution. The teacher is initialized from the same base model as the student but remains fixed throughout training. In our setting, the teacher distribution p_T is defined in Eq. 13, with refusal-oriented conditioning applied only to harmful prompts. The resulting per-rollout OPSD loss is

$$\mathcal{L}_{\text{SD}}(\theta; \mathbf{x}, \mathbf{y}) = \sum_{t=1}^{|\mathbf{y}|} \text{KL}\left(p_\theta(\cdot \mid \mathbf{x}, \mathbf{y}_{<t}) \,\|\, p_T(\cdot \mid \mathbf{x}, \mathbf{y}_{<t})\right), \tag{32}$$

where $\mathbf{y} \sim p_\theta(\cdot \mid \mathbf{x})$ is an on-policy rollout and each KL divergence is computed over the full vocabulary. The training objective averages this loss over sampled prompts and rollouts.

F.5 Experimental Setup

For the GRPO baseline, we utilize the TRL [35] library integrated with vLLM [14] for efficient online rollout generation. We generate $G = 8$ rollouts per prompt to estimate the group-relative advantage. The D_{KL} coefficient is fixed at $\beta = 0.04$ and the clipping parameter ϵ is set to 0.2. To ensure a consistent and fair comparison with THINKSAFE, all other experimental details including the optimizer, batch size, and hardware configuration are maintained identical to the primary experimental setup described in Sec. 5.

For the OPSD baseline, we implement a custom trainer on top of the TRL [35] base trainer that uses vLLM [14] for faster online generation. The student first generates on-policy rollouts, and we then optimize it with a full-vocabulary reverse KL loss against the teacher distribution along the generated trajectories. Due to the substantial latency of repeated rollout generation and teacher scoring, we set the maximum output length to 4096 tokens. This budget is sufficiently large for our setting, as most generated responses fall within this range, as shown in Fig. 9. All remaining optimization and hardware configurations are kept identical the primary experimental setup.

G Limitations

Dependence on latent safety knowledge. Our method fundamentally relies on the hypothesis that the student model retains latent knowledge to identify harm, which refusal steering unlocks. This assumption holds for modern post-trained reasoning models such as Qwen3 and DeepSeek-R1-Distill, which have undergone prior safety alignment. However, for base models that have never been exposed to safety training, the refusal-oriented instruction may fail to elicit valid refusal traces, as there would be little latent safety capability to surface. In such settings, external teacher supervision or RL with safety rewards may remain necessary as an initial step.

Reliance on a safety guard model. THINKSAFE depends on an external safety classifier (Llama-Guard-3) to filter self-generated responses. Although our ablation with WildGuard (Sec. E.4) demonstrates robustness to the specific choice of guard model, the overall quality of the resulting dataset is bounded by the accuracy of the filter. False negatives (unsafe responses labeled as safe) could propagate subtle misalignments into the student, while false positives may discard valid training signals. Advances in safety classifiers would directly benefit our framework.

Scale of evaluation. Our experiments cover model sizes from 0.6B to 8B parameters across two model families. While these cover a representative range of open-source reasoning models, the behavior of THINKSAFE at larger scales (*e.g.*, 70B+) or on closed frontier models remains to be validated. The filtering ratios in Table 3 suggest that larger models retain more of the generated data, indicating that the method may become increasingly effective at scale, but confirming this requires further study.

LoRA-based fine-tuning. Following prior work, we adopt LoRA rather than full parameter fine-tuning. While this choice helps preserve the model's intrinsic capabilities, it also limits the scope of our conclusions regarding distributional discrepancy. Full fine-tuning may exhibit different trade-offs, and we leave a comprehensive comparison to future work.

Single-turn refusal. We focus on single-turn harmful prompts and do not explicitly address multi-turn jailbreaks, adversarial prefixes embedded in long contexts, or agentic settings where safety must be maintained across tool calls. Extending refusal steering to these scenarios is an important direction.

Static offline dataset. We approximate the idealized on-policy objective with a static offline dataset generated once from the initial student. As the student's distribution shifts during fine-tuning, the generated data becomes increasingly off-policy. Iterative self-training, where refusal steering is re-applied to the updated student, may further close this gap at additional computational cost.

H Broader Impacts

Positive impacts. THINKSAFE contributes to the responsible deployment of large reasoning models by mitigating the well-documented safety regression that accompanies reasoning-oriented post-training. By eliminating the dependence on external teacher models, our method lowers the barrier to safety alignment for practitioners who may not have access to large proprietary teachers, and enables smaller research groups and open-source communities to realign reasoning models efficiently. The order-of-magnitude reduction in compute relative to online RL (GRPO) also carries environmental benefits by reducing the energy cost of safety training. More broadly, our theoretical analysis clarifies why self-generated data can preserve native capabilities better than teacher distillation, which may inform future alignment research beyond the safety domain.

Potential negative impacts and mitigation. Research on safety alignment inherently involves working with harmful prompts and jailbreak datasets. We rely on publicly available benchmarks (SafeChain, HarmBench, StrongReject, WildJailbreak, XSTest) that have been vetted by the research community, and we do not generate or release new harmful prompts. The self-generated responses we release consist of refusals and benign reasoning traces, which carry minimal misuse risk. Nevertheless, we acknowledge several concerns:

Dual-use of the method. In principle, the refusal-steering mechanism could be inverted (*e.g.*, with a compliance-oriented instruction) to generate unsafe training data [16]. We note that such attacks are strictly easier without our framework, since reasoning models already comply with many harmful prompts by default, and thus THINKSAFE does not meaningfully expand the attacker's capabilities. Practitioners releasing aligned models should continue to apply standard safeguards such as input and output filtering at deployment.

Over-refusal. Any safety training carries the risk of inducing exaggerated refusal on benign prompts. We explicitly monitor this with XSTest and find that THINKSAFE maintains low over-refusal rates (often below the initial model's), but practitioners should continue to evaluate deployed models on their target distribution of benign queries.

False sense of security. Our method reduces but does not eliminate harmful responses. Models aligned with THINKSAFE should not be treated as adversarially robust, and deployment in high-stakes settings should involve additional safeguards, red-teaming, and monitoring.

Responsible release. We release our code, self-generated datasets, and model adapters under research-oriented licenses, with the aim of enabling reproducibility and further safety research. All released artifacts consist of safe refusals and benign reasoning traces filtered by Llama-Guard-3, and we do not release any model checkpoints that could amplify harm beyond their base models.

I LLM Usage

LLMs as research subjects. The core methodology of this work involves large language models as the primary objects of study. Specifically, we use the Qwen3 family (0.6B, 1.7B, 4B, 8B) and the DeepSeek-R1-Distill family (1.5B, 7B, 8B) as student models, Llama-Guard-3-8B as the safety guard classifier φ, and WildGuard as an alternative classifier for our ablation. All of these models are publicly released and used in accordance with their respective licenses. No proprietary API-based LLMs were used to generate training data, labels, or evaluation judgments.

LLMs for data generation. The self-generated datasets used for THINKSAFE are produced by the student models themselves through refusal steering, as described in Sec. 3. No external teacher LLMs were used in the production of our training data, which is a central design choice of our framework.

LLMs for writing and editing. We used general-purpose LLM assistants for minor writing support, including grammar correction, LaTeX formatting, and rewording for clarity. All technical content, experimental design, theoretical analysis, proofs, and claims were authored and verified by the authors. LLMs were not used to generate research ideas, derive theoretical results, design experiments, produce figures, or draft substantive portions of the paper.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction (Sec. 1) state our three main contributions: (i) formalizing safety realignment as KL projection onto the safe simplex, (ii) proposing THINKSAFE with refusal steering, and (iii) demonstrating superior safety-reasoning trade-offs across Qwen3 and DeepSeek-R1-Distill families. These claims are directly supported by our theoretical analysis in Sec. 4 and the experimental results in Sec. 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We include a dedicated limitations section (Sec. G) discussing dependence on latent safety knowledge, reliance on a safety guard model, scale of evaluation (0.6B to 8B), use of LoRA rather than full fine-tuning, single-turn refusal scope, and the static offline dataset assumption.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results (Lemma 4.1 and Theorem 4.1) are stated with explicit assumptions in Sec. 4, and full proofs are provided in Sec. A.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sec. 5 details the implementation (LoRA configuration, optimizer, learning rate, schedule, batch size, epochs, hardware), datasets, and evaluation protocols. Baseline configurations are provided in Sec. D, sampling details in the appendix, and GRPO details in Sec. F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: As stated in Sec. H, we release our code, self-generated datasets, and model adapters under research-oriented licenses. All underlying prompts come from publicly available datasets (SafeChain, HarmBench, StrongReject, WildJailbreak, XSTest).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training details (AdamW, learning rate 1×10^{-5}, cosine scheduler with 10% warmup, batch size 8, 3 epochs, LoRA rank 32, $\alpha = 16$, dropout 0.05) and evaluation details (sampling temperature, top-p, top-k, max tokens, pass@1 over 8 samples) are given in Sec. 5 and Sec. C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: We report average pass@1 over 8 sampled responses per prompt for reasoning benchmarks, but do not report formal error bars across training runs due to the computational cost of fine-tuning multiple models at each scale. We note that safety evaluation uses a single response per prompt following standard protocol.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report that all experiments were run on 2 NVIDIA H100 GPUs (Sec. 5), and the training time comparison with GRPO (including wall-clock hours) is reported in Fig. 3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research conforms with the NeurIPS Code of Ethics. We use only publicly available datasets vetted by the research community and focus on improving safety alignment rather than producing harmful content, as discussed in Sec. H.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Sec. H discusses positive impacts (lowering barriers to safety alignment, reducing compute costs) and potential negative impacts (dual-use, over-refusal, false sense of security) along with mitigation strategies.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: As described in Sec. H, our released artifacts consist exclusively of safe refusals and benign reasoning traces filtered by Llama-Guard-3. We do not release harmful prompts or model checkpoints that amplify harm beyond their base models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All models (Qwen3, DeepSeek-R1-Distill, Llama-Guard-3, WildGuard) and datasets (SafeChain, HarmBench, StrongReject, WildJailbreak, XSTest, GSM8K, MATH500, AIME24, GPQA) are cited with their original papers and used in accordance with their respective licenses, as noted in Sec. I and throughout the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The self-generated datasets and LoRA adapters released alongside the paper are documented in Sec. B.2 (dataset statistics), Sec. 5 (training configurations), and Sec. H (release and licensing).

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve research with human subjects, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are the central research subjects of this work (Qwen3 and DeepSeek-R1-Distill families as students; Llama-Guard-3 and WildGuard as safety classifiers). Their usage is fully described in Sec. I, along with the clarification that LLMs were used only for minor writing assistance and not for generating research ideas, experiments, proofs, or training data.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.