

Beyond SFT: Reinforcement Learning for Safer Large Reasoning Models with Better Reasoning Ability

Anonymous ACL submission

Abstract

Large reasoning models (LRMs) extend large language models by generating explicit chain-of-thought (CoT) reasoning, significantly improving mathematical and logical problem solving. However, this explicit reasoning process also introduces new safety risks, as unsafe behaviors often emerge within intermediate reasoning trajectories, even when final answers appear harmless. Existing safety alignment approaches primarily rely on supervised fine-tuning (SFT) over safety-oriented long CoT datasets. While intuitive, we find that SFT produces inconsistent safety improvements, degrades reasoning ability, and generalizes poorly across model families. These limitations suggest that purely supervised approaches are insufficient for robust safety alignment in LRMs. To address this, we investigate reinforcement learning (RL) as a complementary optimization framework for LRM safety training. Unlike SFT, RL directly optimizes model policies with reward feedback, enabling more adaptive and stable alignment. Extensive experiments across multiple model families and benchmarks show that RL achieves stronger and more consistent safety gains while maintaining reasoning competence. Further analysis of reflection dynamics and token-level entropy reveals that RL suppresses unsafe exploratory reasoning while preserving reflective depth, leading to safer and more reliable reasoning processes.

1 Introduction

Large reasoning models (LRMs) extend the capability of large language models (LLMs) by incorporating explicit chain-of-thought (CoT) reasoning (Wei et al., 2022; Jaech et al., 2024; Guo et al., 2025). By generating intermediate reasoning steps before producing final answers, LRMs demonstrate remarkable improvements in mathematical, logical, and scientific problem solving. However, this explicit reasoning process introduces new safety challenges that differ fundamentally from those of

conventional LLMs. Recent studies (Wang et al., 2025a; Zhou et al., 2025; Fang et al., 2025) reveal that unsafe behaviors often arise within reasoning trajectories of LRMs, such as toxic or deceptive intermediate thoughts, even when the final answers appear benign. Moreover, enhanced reasoning capabilities can amplify harmful behaviors relative to standard LLMs (Zhou et al., 2025), meaning that stronger reasoning does not necessarily imply safer behavior. This phenomenon, known as *unsafe reasoning*, highlights that safety risks in LRMs arise from the reasoning process itself rather than surface-level outputs. We further find that mixture-of-thinking models also exhibit lower safety when the thinking mode is enabled compared to the non-thinking mode.

To mitigate the safety risks, prior works have primarily relied on supervised fine-tuning (SFT) on safety-oriented long CoT datasets such as STAIR (Zhang et al., 2025c), STAR-1 (Wang et al., 2025c), and SafeChain (Jiang et al., 2025). SFT aims to align model behaviors by directly supervising safe reasoning trajectories or refusal patterns. While intuitive and widely adopted, the effectiveness of SFT on LRMs has not been systematically examined. In this work, we show that SFT-based safety enhancement has notable limitations and is less effective than commonly believed. We ask:

(Q) What are the limitations and underlying causes of SFT for LRM safety enhancement, and how can it be improved?

To address (Q), we conduct a comprehensive evaluation across multiple SFT datasets and model families, revealing several key limitations. First, safety gains from SFT fail to generalize across architectures; for instance, datasets distilled from DeepSeek-based models transfer poorly to structurally different ones such as GRANITE-4.0-TINY-PREVIEW (IBM, 2025). Second, SFT performance is susceptible to dataset quality and model compat-

082 ability; Training on noisy or mismatched data often leads to unstable or even negative safety gains. 083
084 Third, excessive SFT can impair reasoning ability, as over-regularized models lose flexibility and 085
086 suffer from catastrophic forgetting (Shenfeld et al., 2025). These findings indicate that purely supervised 087
088 approaches are insufficient for achieving robust and scalable safety alignment in LRMs. 089

090 To address these limitations, we investigate reinforcement learning (RL) as a complementary optimization 091
092 paradigm for LRM safety alignment. Unlike SFT, which relies on fixed supervision signals, RL optimizes 093
094 model policies directly toward safety-aligned objectives through reward feedback. Despite RL’s popularity 095
096 in enhancing the reasoning capabilities of LLMs (Guo et al., 2025; Jaech et al., 2024), its effectiveness 097
098 for safety training in LRMs remains largely unexplored in the literature. In this study, we leverage 099
100 RL-based optimization as a general framework to enhance safety while preserving reasoning capability. We 101
102 summarize our **contributions** below: 103

- 104 • We find that mixture-of-thinking models exhibit reduced safety in the thinking mode compared to 105
106 the non-thinking mode. Furthermore, our comprehensive empirical study on SFT for LRM safety 107
108 alignment reveals several key limitations, including weak cross-model generalization, inconsistent 109
110 safety across datasets, and noticeable reasoning degradation. 111
- 112 • We employ a RL-based alignment framework to achieve consistent improvements in both safety and 113
114 reasoning. 115
- 116 • We conduct extensive experiments across diverse benchmarks and model families, highlighting the 117
118 limitations of SFT and the advantages of RL. Fine-grained analyses of reflection dynamics and token- 119
120 level entropy reveal that RL enhances safety by suppressing unsafe exploratory reasoning while 121
122 preserving reflective depth in legitimate reasoning tasks. 123

124 2 Related Work

125 **Safety risks in LRMs.** LRMs introduce distinct safety vulnerabilities due to their explicit chain- 126
127 of-thought (CoT) reasoning traces (Wang et al., 2025a). Unlike standard LLMs, their intermediate 128
129 reasoning steps can amplify unsafe behaviors such as toxicity, deception, or harmful compliance, 130
131 even when final answers appear benign (Zhou et al., 2025; Arrieta et al., 2025a,b). Beyond this, re-

132 searchers have identified additional risk categories, including agentic misbehavior (Xu et al., 2025; 133
134 Barkur et al., 2025; He et al., 2025), multilingual disparities (Ying et al., 2025; Zhang et al., 135
136 2025a), and multimodal vulnerabilities (Fang et al., 2025). Reasoning-level adversarial attacks, such 137
138 as BadChain (Xiang et al., 2024), DarkMind (Guo and Tourani, 2025), and Shadow-CoT (Zhao et al., 139
140 2025), further expose how unsafe exploration paths in CoT generation lead to harmful outputs. Overall, 141
142 safety failures in LRMs stem primarily from unsafe reasoning dynamics rather than surface responses. 143
144 In this work, we focus on *harmful request compliance*, the most practically relevant and frequently 145
146 observed form of unsafe reasoning in real-world settings. 147

Defenses for LRMs. Recent studies defend LRMs from unsafe reasoning through both training- 148
149 and inference-level approaches (Wang et al., 2025a). Training-based defenses mainly rely on 150
151 supervised fine-tuning (SFT), preference optimization, or reinforcement learning (RL). SFT-based 152
153 methods such as STAR-1 (Wang et al., 2025c), SafeChain (Jiang et al., 2025), STAIR (Zhang et al., 154
155 2025c), and RealSafe-R1 (Zhang et al., 2025b) align reasoning via long CoT datasets with safety 156
157 annotations. R2D (Zhu et al., 2025) formulates safety as contrastive preference optimization using 158
159 safe–unsafe reasoning pairs, and is also reused as an SFT dataset bridging instruction- and reward- 160
161 based paradigms. Offline RL methods, including DPO in STAIR (Zhang et al., 2025c) and 162
163 SaRO (Mou et al., 2025), optimize pre-collected trajectories, while Deliberative Alignment (Guan 164
165 et al., 2024) improves final-answer safety through policy-guided reasoning. In contrast, our *online* 166
167 RL formulation enables better generalization and lower dependence on curated supervision. 168
169

170 Inference-time defenses such as ZeroThink/Less-Think/MoreThink (Jiang et al., 2025) and Think- 171
172 ing Intervention (Wu et al., 2025) dynamically guide reasoning to prevent unsafe exploration. Al- 173
174 though effective, these approaches require continuous token-level monitoring or auxiliary controllers, 175
176 leading to high cost and limited scalability compared with training-time reasoning alignment. 177

178 3 Revisiting and Extending the Safety Analysis of LRMs

180 In this section, we re-examine the safety of LRMs and show that, at comparable scale, they are sub- 181

stantially less safe than standard LLMs. We then analyze the root causes of these vulnerabilities and apply a fine-grained evaluation framework to pinpoint which components of generated responses exhibit unsafe behavior. Finally, we extend our study to mixture-of-thinking models, revealing that they too suffer from safety weaknesses, particularly when the explicit reasoning mode is activated.

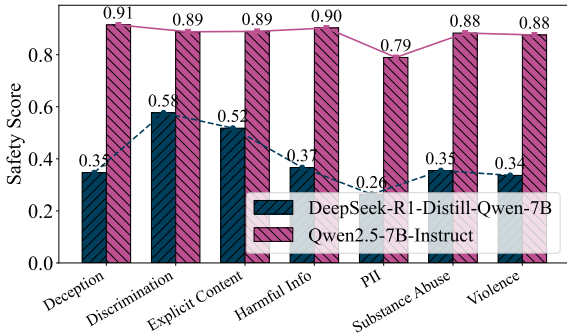


Figure 1: Safety performance on the AttaQ benchmark, comparing the LRM DEEPSEEK-R1-DISTILL-QWEN-7B with the standard instruction-tuned model QWEN2.5-7B-INSTRUCT across multiple safety categories. Higher scores indicate better safety. The LRM consistently lags behind the LLM of the same scale, revealing a pronounced safety gap.

Safety gap between LRMs and standard LLMs.

Recent studies (Wang et al., 2025c; Zhang et al., 2025b; Jiang et al., 2025) have revealed that LRMs, despite their strong performance on reasoning-intensive tasks such as mathematics and programming, exhibit pronounced safety vulnerabilities. To further investigate this phenomenon, we evaluate safety performance using the AttaQ benchmark (Kour et al., 2023), a comprehensive suite of adversarial questions designed to test the harmlessness of language models. Specifically, we compare the LRM DEEPSEEK-R1-DISTILL-QWEN-7B with its standard instruction-tuned counterpart QWEN2.5-7B-INSTRUCT, which share the same architecture and scale.

As shown in **Figure 1**, the LRM consistently underperforms across all safety categories of AttaQ. For instance, in the *Deception* category, DEEPSEEK-R1-DISTILL-QWEN-7B achieves a score of only 0.35 compared to 0.91 for QWEN2.5-7B-INSTRUCT. A similar disparity is observed in *Harmful Information* (0.37 vs. 0.90). Even in categories where the LRM fares relatively better, such as *Explicit Content*, its score (0.52) still lags far behind the baseline (0.89). These results underscore a pronounced **safety gap**: while standard LLMs such as QWEN2.5-7B-INSTRUCT maintain high and stable safety scores across categories, LRMs

of the same size exhibit substantially greater risks. This gap raises a critical question: *what underlying mechanisms drive the unsafe behaviors characteristic of LRMs?*

Granular safety evaluation for LRMs. To better understand where unsafe behaviors originate in LRMs, we perform a fine-grained safety analysis that separates model outputs into the *final answer* (y) and the *whole response* ($t + y$), where t denotes the reasoning trajectory. Following the design of reasoning models such as OpenAI-O1 (Jaech et al., 2024), DEEPSEEK-R1 (Guo et al., 2025), and GEMINI-2.5 (Comanici et al., 2025), each output is represented as (t, y) . We compare the safety of the final answer and the whole response to identify whether unsafe behaviors emerge during reasoning.

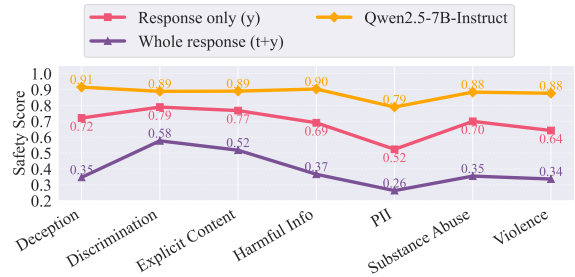


Figure 2: Safety performance on AttaQ for different response components of DEEPSEEK-R1-DISTILL-QWEN-7B (*response y* and *whole response t + y*), compared with QWEN2.5-7B-INSTRUCT. Higher scores indicate better safety.

As shown in **Figure 2**, the whole-response safety remains markedly lower than that of the final answers, consistent with prior observations of reasoning-level degradation in LRMs (Zhou et al., 2025; Fang et al., 2025). This result suggests that unsafe behaviors are primarily introduced during the reasoning process and carried into the overall response, underscoring the need for alignment methods that explicitly regulate internal reasoning rather than relying solely on output-level control.

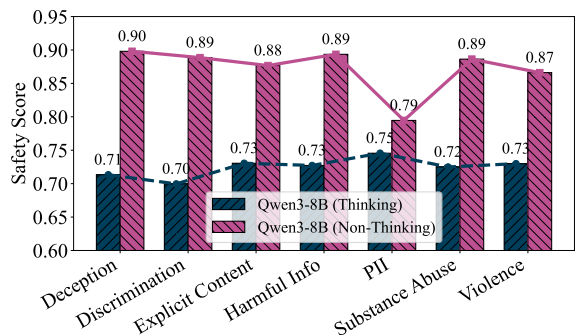


Figure 3: Safety performance of QWEN3-8B on the AttaQ benchmark in *non-thinking* and *thinking* modes. Higher scores indicate stronger safety. All other settings follow Figure 1.

Safety of mixture-of-thinking models. The above analysis highlights that unsafe behaviors in LRMs primarily stem from their reasoning trajectories. A natural question is whether these vulnerabilities persist in more recent *mixture-of-thinking* models, which can flexibly switch between reasoning and non-reasoning modes. For example, QWEN3 (Team, 2025) alternates between a *thinking mode*, resembling LRM-style generation, and a *non-thinking mode*, resembling standard LLM behavior. Figure 3 shows the safety of QWEN3-8B on the AttaQ benchmark under both thinking and non-thinking modes. The results reveal a clear degradation in safety when the thinking mode is enabled. For instance, in the *Deception* category, the safety score drops by 0.19 (0.71 vs. 0.90). This suggests that mixture-of-thinking models inherit the same vulnerabilities as LRMs, as reasoning-style generation implicitly amplifies unsafe behaviors. These findings again underscore the need for alignment strategies that explicitly regulate not only the final response but also the internal reasoning process.

4 From Supervised to RL-based Safety Fine-Tuning: Limitations and Promises

In this section, we revisit the limitations of supervised fine-tuning when applied to LRMs for safety alignment. We first find that the effectiveness of SFT is strongly dependent on model type: its safety gains do not transfer reliably across model families, and are often significantly weaker on architectures or pretraining distributions that differ from those of its training source, such as GRANITE-4.0-TINY-PREVIEW. Moreover, SFT performance strongly depends on dataset quality: while STAR-1 (Wang et al., 2025c) improves safety, other long chain-of-thought datasets such as SafeChain (Jiang et al., 2025) and R2D-R1 (Zhu et al., 2025) provide limited or even negative effects, highlighting SFT’s sensitivity to data alignment and noise. Even when safety improves, SFT frequently reduces reasoning ability on benchmarks like GPQA-Diamond and AIME24/25, revealing an inherent trade-off between safety enhancement and reasoning preservation. Thus, SFT alone is insufficient for robust LRM safety alignment. By contrast, we find that RL provides a more general and data-efficient alternative that achieves better safety–reasoning balance without requiring high-quality reference data.

Poor transferability of SFT across model families. We first identify a key limitation of SFT: its sensitivity to model architecture and pretraining distribution. While SFT on STAR-1 (Wang et al., 2025c) has shown substantial safety improvements on DEEPSEEK-R1–based distilled models, its effectiveness has rarely been tested on models with different architectures or pretraining corpora. Figure 4 compares the average safety scores of STAR-1–trained models on DEEPSEEK-R1-DISTILL-QWEN-7B and GRANITE-4.0-TINY-PREVIEW using the AttaQ benchmark. Notably, GRANITE-4.0-TINY-PREVIEW, a hybrid Mamba–Transformer model (Gu and Dao, 2023), exhibits much smaller safety gains after SFT (0.02 vs. 0.39). Although GRANITE-4.0-TINY-PREVIEW starts with a higher baseline safety score, post-SFT the DeepSeek-based model surpasses it. The sharp contrast between the negligible safety gain of SFT on GRANITE-4.0-TINY-PREVIEW and its notable improvement on DEEPSEEK-R1-DISTILL-QWEN-7B highlights SFT’s sensitivity to the choice of base model and its limited transferability, even when effective on a specific model.

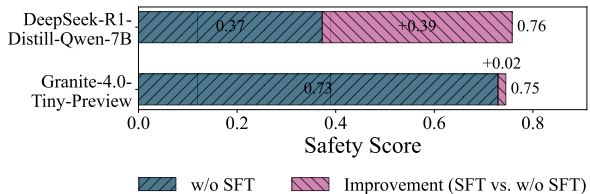


Figure 4: Average safety scores on the AttaQ benchmark for GRANITE-4.0-TINY-PREVIEW and DEEPSEEK-R1-DISTILL-QWEN-7B, before and after SFT with STAR-1 data. Bars show baseline performance without SFT (blue) and the improvement after SFT (pink).

We hypothesize that the above discrepancy arises because STAR-1 was distilled from DeepSeek-R1, making its data distribution and reasoning format naturally compatible with DeepSeek-derived models. In contrast, the distinct pretraining distribution of GRANITE-4.0-TINY-PREVIEW limits its ability to leverage STAR-1 effectively. To verify this hypothesis, we employ the Min-K% Probability (Min-K% Prob) metric (Shi et al., 2023) to quantify model memorization by computing

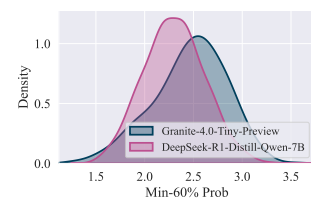


Figure 5: Distribution of Min-K% Probability (Min-60% Prob) (Shi et al., 2023) values for DEEPSEEK-R1-DISTILL-QWEN-7B and GRANITE-4.0-TINY-PREVIEW on STAR-1 data. Lower scores indicate stronger memorization.

the average *negative* log-likelihood over the lowest $K\%$ predicted tokens in each sequence; lower values indicate stronger memorization. **Figure 5** reports results for $K = 60\%$, showing that DEEPSEEK-R1-DISTILL-QWEN-7B consistently yields lower Min- $K\%$ Prob scores than GRANITE-4.0-TINY-PREVIEW on STAR-1 data. This confirms that the STAR-1 data distribution is more aligned with DeepSeek-derived models, explaining their larger safety gains of SFT on DEEPSEEK-R1-DISTILL-QWEN-7B.

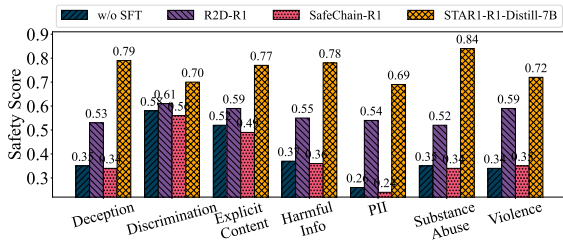


Figure 6: Comparison of SFT-trained models on AttaQ. The base model is DEEPSEEK-R1-DISTILL-QWEN-7B.

Inconsistent safety gains from SFT across datasets. Prior studies (Wang et al., 2025c; Jiang et al., 2025; Zhang et al., 2025b,c) show that safety-oriented long chain-of-thought training can improve LRM safety. However, few works have systematically compared SFT performance across different datasets under a unified evaluation setup. **Figure 6** shows the performance of SFT trained on different datasets, including STAR-1 (Wang et al., 2025c), R2D-R1 (Zhu et al., 2025), and SafeChain (Jiang et al., 2025), evaluated on the AttaQ benchmark using DEEPSEEK-R1-DISTILL-QWEN-7B as the base model. Only STAR-1 yields consistent safety improvements across harm categories, while others provide marginal or even negative gains. For example, the SafeChain-trained model drops from 0.52 to 0.49 in the *Explicit Content* category, suggesting that SafeChain’s lower quality or less strictly filtered data introduces noisy supervision compared to STAR-1. Moreover, as shown later, this degradation often extends to reasoning performance. Thus, the safety benefits of SFT remain highly dataset-dependent, and suboptimal data can undermine both safety and reasoning.

Impact of SFT on reasoning performance. Beyond SFT’s data–model sensitivity in achieving safety gains, prior work (Wang et al., 2025c) also reveals a trade-off between safety and reasoning performance under SFT. To re-examine this effect, we evaluate the safety-enhanced models by SFT in Fig. A1, across three reasoning

benchmarks: AIME24/25 (MAA Committees), MATH500 (Lightman et al., 2023), and GPQA-Diamond (Rein et al., 2024). It shows that while SFT on STAR-1 largely preserves reasoning ability on AIME24/25, it exhibits a noticeable drop on GPQA-Diamond. Models trained on SafeChain or R2D-R1 perform even worse, with reasoning scores far below the original LRM. These results confirm that SFT’s safety gains often come at the cost of reasoning degradation, and excessive alignment further amplifies this trade-off. This reflects SFT’s vulnerability to catastrophic forgetting during additional training (Shenfeld et al., 2025).

The aforementioned limitations of SFT motivate us to ask: *Is there a general approach that can improve the safety of LRMs without relying on strong data-model prerequisites, while simultaneously preserving their reasoning capabilities?*

RL to jointly optimize safety and reasoning. To overcome the limitations of SFT, we employ RL as an alternative, which has been shown to enhance generalization (Kirk et al., 2023; Chu et al., 2025) and mitigate catastrophic forgetting (Shenfeld et al., 2025). However, despite its widespread use for enhancing reasoning, the effectiveness of RL for safety training in LRMs remains largely unexplored in the literature.

In the RL paradigm, an LRM is treated as a policy π_θ that generates a full response (\mathbf{t}, \mathbf{y}) , consisting of both the reasoning trajectories \mathbf{t} and the final answer \mathbf{y} , given a prompt \mathbf{x} . The training objective is to maximize the expected reward, or equivalently, minimize the negative expected reward:

$$\mathcal{L}_{\text{RL}}(\theta) = -\mathbb{E}_{(\mathbf{t}, \mathbf{y}) \sim \pi_\theta(\cdot | \mathbf{x})} [R(\mathbf{x}, \mathbf{t} + \mathbf{y})], \quad (1)$$

where θ represents the LRM, and $R(\mathbf{x}, \mathbf{t} + \mathbf{y})$ encodes safety criteria. In our implementation, we adopt REINFORCE++ (Hu et al., 2025), an enhanced policy-gradient algorithm that removes the need for a critic network while incorporating several stability techniques, including token-level KL divergency penalties, proximal policy optimization (PPO)-style clipping, mini-batch updates, and normalized advantage estimation. These modifications make Reinforce++ both simpler and more efficient than PPO (Schulman et al., 2017), while retaining stable optimization dynamics. We use this approach to explore whether RL is a more effective framework for improving the safety of LRMs without degrading their reasoning performance.

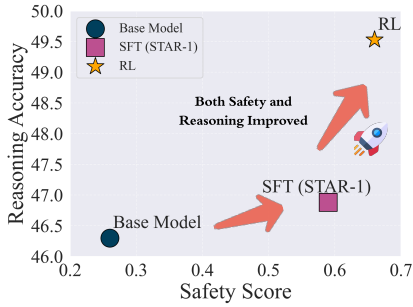


Figure 7: Safety and reasoning trade-off visualization for DEEPSEEK-R1-DISTILL-QWEN-7B under different alignment strategies. The x-axis is the safety score on AIR-Bench, and the y-axis denotes accuracy on AIME24.

To validate the effectiveness of our RL formulation, we conduct an initial comparison among the base model, the SFT baseline (trained on STAR-1), and the RL-aligned model (also trained on STAR-1) on two representative benchmarks: AIR-Bench (Zeng et al., 2024) for safety and AIME24 for reasoning. As shown in **Figure 7**, SFT substantially improves safety alignment while maintaining comparable reasoning accuracy. In contrast, our RL-aligned model achieves simultaneous gains in both dimensions, indicating that reinforcement optimization can enhance safety without compromising, and even slightly improving reasoning capability.

5 Experiments

5.1 Experimental Setup

Evaluation benchmarks. To evaluate our approach, we assess both safety and reasoning performance. For safety evaluation, we consider a diverse set of benchmarks: **AttaQ** (Kour et al., 2023), which tests adversarial safety across seven harm categories (*e.g.*, deception, violence, hate) and scores outputs with OPENASSISTANT/REWARD-MODEL-DEBERTA-V3-LARGE-V2 normalized to $[0, 1]$; and **AIR-Bench** (Zeng et al., 2024), a regulation-grounded benchmark covering 314 fine-grained risks across four domains, where GPT-4o automatically judges refusal rates (higher is safer). In addition, we include StrongReject (Souly et al., 2024), JBB-Behaviors (Chao et al., 2024), and WildChat (Zhao et al., 2024) to evaluate the model’s ability to refuse disallowed content, as well as WildJailbreak (Jiang et al., 2024) to assess robustness to adversarial jailbreak attacks. All additional benchmarks follow the evaluation protocol of STAR-1 (Wang et al., 2025c), using stochastic decoding with top-k sampling ($k = 50$) at temperature 1.0 and a maximum generation length of 9,000 tokens.

For reasoning, we use **MATH500** (Lightman et al., 2023), **AIME24/25** (MAA Committees), and **GPQA-Diamond** (Rein et al., 2024), covering multi-step mathematics and graduate-level STEM problems. All reasoning evaluations follow Guo et al. (2025) with a 32k-token limit, nucleus sampling (temperature 0.6, top- $p = 0.95$).

Models and datasets. To evaluate the effectiveness of RL in enhancing both safety and reasoning, we conduct experiments on four representative model families: (i) DEEPSEEK-R1-DISTILL-QWEN-7B, a distilled large reasoning model (LRM); (ii) QWEN3-8B, a recent LRM capable of operating in both *thinking* and *non-thinking* modes; (iii) GRANITE-4.0-TINY-PREVIEW, which adopts a hybrid Mamba-Transformer architecture; and (iv) DEEPSEEK-R1-DISTILL-LLAMA-8B, a distilled LRM based on a non-Qwen architecture. For supervised fine-tuning (SFT) baselines, we fine-tune models on several widely used safety-oriented long chain-of-thought (CoT) datasets, including STAR-1 (Wang et al., 2025c), R2D-R1 (Zhu et al., 2025), and SafeChain (Jiang et al., 2025). We also incorporate the **contrastive pivot optimization (CPO)** approach from R2D (Zhu et al., 2025), trained on the R2D-R1 dataset, as a preference-optimization baseline. For RL training, we use the same prompt distribution as STAR-1 (Wang et al., 2025c) to ensure a fair comparison, isolating the effect of the training paradigm rather than differences in data.

Implementation details. For SFT, we use the OpenRLHF framework (Hu et al., 2024), following the training configuration recommended by STAR-1 (Wang et al., 2025c). Models are trained for 5 epochs with a learning rate of 5×10^{-6} and a batch size of 256. For SafeChain and R2D-R1, we train for only 1 epoch, as we observed that longer training substantially degrades reasoning performance, and both datasets are relatively large; other hyperparameters remain the same.

For RL, we adopt REINFORCE++ as the training algorithm and use SKYWORK-REWARD-V2-LLAMA-3.1-8B (Liu et al., 2025) as the reward model, which achieves state-of-the-art performance on RewardBench (Malik et al., 2025). The reward model provides token-level reward signals for each generated response. We train for 500 episodes across the same prompt distribution, ensuring a fair comparison with SFT baselines.

Table 1: Performance comparison of supervised fine-tuning (SFT), preference optimization (CPO), and RL methods on safety and reasoning benchmarks across two large reasoning model (LRM) families. **Avg. Safety** denotes the arithmetic mean across all safety benchmarks. All metrics are higher-is-better, and the best result in each column is highlighted in **bold**.

Model	Safety						Reasoning				
	AttaQ	AIR-Bench	Strong Reject	JBB	WildChat	Wild Jailbreak	Avg. Safety	GPQA -Diamond	MATH500	AIME24	AIME25
DeepSeek-R1 distilled family											
DeepSeek-R1-Distill-Qwen-7B	0.37	0.26	0.58	0.52	0.75	0.52	0.50	49.24	92.00	46.30	30.52
+ SFT (R2D-R1)	0.56	0.41	0.72	0.66	0.80	0.74	0.65	46.53	86.80	39.64	29.38
+ SFT (SafeChain)	0.37	0.25	0.63	0.58	0.86	0.65	0.56	48.48	91.05	42.60	28.64
+ SFT (STAR-1)	0.76	0.59	0.99	0.98	0.85	0.85	0.84	47.54	91.80	46.88	31.87
+ CPO	0.59	0.41	0.69	0.66	0.79	0.72	0.64	47.85	90.75	41.67	27.86
+ Ours	0.78	0.66	1.00	1.00	0.95	0.98	0.90	49.68	92.30	49.53	32.14
Qwen3 family											
Qwen3-8B (thinking)	0.73	0.40	0.94	0.96	0.64	0.62	0.72	59.53	96.40	74.22	40.57
+ SFT (R2D-R1)	0.75	0.43	0.87	0.83	0.88	0.88	0.77	51.83	91.70	60.05	36.04
+ SFT (SafeChain)	0.49	0.29	0.84	0.76	0.87	0.83	0.68	57.01	95.00	67.60	39.06
+ SFT (STAR-1)	0.78	0.51	0.92	0.94	0.74	0.83	0.79	59.79	95.50	74.69	42.55
+ CPO	0.79	0.55	0.88	0.88	0.82	0.90	0.80	53.54	95.55	68.33	42.29
+ Ours	0.81	0.58	1.00	1.00	0.84	0.88	0.85	58.33	96.40	75.16	44.11

5.2 Experimental Results

RL enhances safety while preserving reasoning. In Table 1, we evaluate the safety and reasoning performance of RL methods on AttaQ, AIR-Bench, and several reasoning benchmarks, including MATH500, AIME24, AIME25, and GPQA-Diamond. We compare ours against multiple baselines across two representative model families: DEEPSEEK-R1-DISTILL-QWEN-7B, a distilled large reasoning model derived from DEEPSEEK-R1, and QWEN3-8B, a recent model that supports both *thinking* and *non-thinking* modes. The baselines include SFT models fine-tuned on widely used safety-oriented long chain-of-thought datasets, STAR-1, R2D-R1, and SafeChain, as well as a preference-based contrastive pivot optimization (CPO) method from R2D.

First, ours achieves the best safety performance compared with all baselines and the base models. For example, on DEEPSEEK-R1-DISTILL-QWEN-7B, STAR-1 training achieves an average safety score of 0.85, whereas our method, trained on the same STAR-1 prompts, attains a score of 0.95, an absolute improvement of 0.10. These results indicate that RL has greater potential to enhance safety performance, likely due to its superior generalization capability compared to SFT and CPO, consistent with prior observations in Kirk et al. (2023).

Second, ours also outperforms SFT on reasoning ability. On DEEPSEEK-R1-DISTILL-QWEN-7B, ours achieves the strongest results across all reasoning benchmarks, with a particularly notable gain on AIME24 (+3.23, 49.53 vs. 46.30 compared to the base model). By contrast, SFT often degrades reasoning performance: for example, STAR-1 training reduces GPQA accuracy from 49.24 (no SFT) to 47.54, whereas ours maintains performance at 49.68, comparable to the baseline. This indicates

that RL better preserves reasoning skills and prior knowledge, mitigating the catastrophic forgetting issues discussed in Shenfeld et al. (2025).

Third, we find that these improvements are consistent across model families. For example, in the QWEN3-8B family, ours achieves the highest safety scores on average safety score 0.85, while also attaining the strongest reasoning results on MATH500 (96.40), AIME24 (75.16), and AIME25 (44.11). Such consistency across both distilled and hybrid LRM architectures underscores the robustness of ours as an effective approach for safety alignment without compromising reasoning.

RL improves safety across AIR-Bench categories. The radar plots in Figure 8 compare the

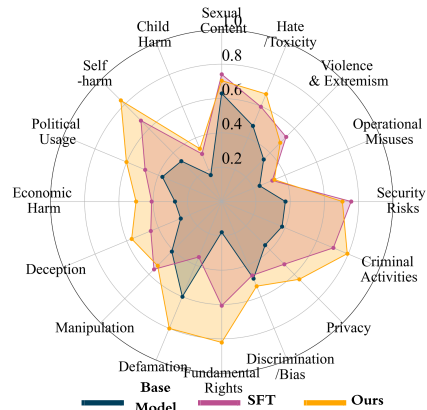


Figure 8: Category-level results on AIR-Bench for QWEN3-8B. This compares the base model (blue), SFT on STAR-1 (purple), and ours (orange). Higher scores indicate safer behavior.

base model, the SFT baseline (STAR-1), and our RL-aligned model for QWEN3-8B, illustrating refusal rates across diverse safety categories such as harmful content, privacy, and fundamental rights. Across nearly all dimensions, RL achieves the highest safety scores. For instance, on QWEN3-8B, RL improves refusal rates in high-risk areas by notable margins, while SFT shows smaller or inconsistent

gains and even underperforms the baseline in certain categories, like *Defamation*. Overall, these results confirm that RL not only enhances overall safety averages (Table 1) but also provides **broad and fine-grained safety improvements across diverse risk types**. In particular, RL substantially reduces unsafe outputs in sensitive domains, like *self-harm*, demonstrating superior robustness and generalizability compared to SFT. Additional radar plots for DEEPSEEK-R1-DISTILL-QWEN-7B are provided in Figure A2, showing consistent trends.

RL improves safety by regulating reflection behavior. To further understand how RL shapes LRM behavior beyond benchmark scores, we move beyond benchmark scores and analyze token-level uncertainty within reflection sequences. Following (Wang et al., 2025b), we compute the next-token entropy at position t as $H_t = -\sum_{j=1}^V p_{t,j} \log p_{t,j}$, where H_t denotes the entropy (in bits), V is the vocabulary size, and $p_{t,j}$ is the model’s predicted probability of token j . High-entropy tokens indicate *forks* in the chain of thought—points where the model explores multiple potential reasoning branches.

We adopt a teacher-forcing protocol: for each prompt, we first extract reflection sequences from the base model (DEEPSEEK-R1-DISTILL-QWEN-7B) on AIME24 (reasoning) and AttaQ (unsafe). We then force the base, SFT, and RL models to follow these same sequences and measure H_t at each reflection token. Table 2 reports the average entropy over common reflection tokens (“wait”, “hmm”, “but”, “alternatively”), on unsafe prompts from AttaQ and reasoning prompts from AIME24, comparing the base model, SFT, and RL variants, with DEEPSEEK-R1-DISTILL-QWEN-7B as the base model.

Table 2: Comparison of average reflection-token entropy following (Wang et al., 2025b) (wait, hmm, but, alternatively) across AttaQ (Safety) and AIME24 (Reasoning). The base model is given by DEEPSEEK-R1-DISTILL-QWEN-7B, and its SFT and RL variants are fine-tuned on STAR-1.

Model	AttaQ	AIME24
Base model	0.24	3.12
SFT	0.12	2.73
RL	0.09	3.00

On AttaQ, RL achieves the lowest reflection entropy (0.09 vs. 0.24 for base and 0.12 for SFT), indicating more decisive behavior and suppressed unsafe exploration. On AIME24, SFT exhibits the lowest entropy (2.73), consistent with its degraded

reasoning performance due to reduced exploration, while RL maintains entropy close to the base model, preserving multi-path reasoning capacity.

Across all variants, reflection entropy is substantially higher on AIME24 than AttaQ, reflecting intrinsic task differences: reasoning tasks require broader exploration, whereas safety prompts benefit from low-uncertainty, early-terminating behavior.

Additional experiments. We further extend our study with several additional analyses to validate the robustness of our conclusions. First, as shown in Table A1, on the DEEPSEEK-R1-DISTILL-LLAMA-8B model, RL consistently outperforms SFT in terms of both safety and reasoning performance, demonstrating a superior safety and reasoning trade-off. Second, Table A2 shows that RL methods equipped with different reward models achieve better performance than SFT across safety and reasoning benchmarks, indicating that the observed gains are not sensitive to a specific reward model choice. Third, under equal training budgets, results in Table A3 reveal that SFT fails to match the performance of RL, further highlighting the effectiveness of RL-based alignment. In addition, we evaluate safety performance on GRANITE-4.0-TINY-PREVIEW (Figure A3), confirming that our findings generalize beyond the Qwen and distilled LRM families. Finally, qualitative examples in Table A4 illustrate that RL produces safer yet coherent reasoning trajectories.

6 Conclusions

In this paper, we present a comprehensive study of safety alignment in Large Reasoning Models (LRMs). Through extensive experiments, we reveal that supervised fine-tuning (SFT) suffers from unstable safety gains, degraded reasoning ability, and poor cross-model generalization. To overcome these limitations, we introduce a reinforcement learning (RL)-based alignment framework that directly optimizes reasoning trajectories via reward feedback. Our results across multiple benchmarks and model families demonstrate that RL achieves consistent improvements in both safety and reasoning performance. Fine-grained analyses of reflection dynamics and reasoning entropy further show that RL enables safer, more controlled reasoning behavior. Overall, this work establishes RL as a scalable and effective paradigm for reasoning-aware safety alignment in LRMs.

7 Limitations

While our study establishes reinforcement learning (RL) as an effective paradigm for reasoning-aware safety alignment, several limitations remain. First, our experiments are conducted on medium-scale LRMs (up to 8B parameters); extending the analysis to larger frontier models (*e.g.*, 70B or above) may reveal different optimization dynamics and scaling behaviors. Second, our evaluation relies primarily on automated benchmarks and model-based metrics; incorporating human evaluation would provide deeper insights into nuanced safety failures and reasoning quality that may not be fully captured by existing benchmarks. Third, although we demonstrate consistent safety and reasoning gains, our approach relies on existing reward models, whose biases and coverage may affect training outcomes. Future work will explore scaling RL-based safety alignment to larger models, integrating adversarial robustness evaluation, and developing more fine-grained reward signals for reasoning-process regulation.

References

Aitor Arrieta, Miriam Ugarte, Pablo Valle, José Antonio Parejo, and Sergio Segura. 2025a. Early external safety testing of openai’s o3-mini: Insights from the pre-deployment evaluation. *arXiv preprint arXiv:2501.17749*.

Aitor Arrieta, Miriam Ugarte, Pablo Valle, José Antonio Parejo, and Sergio Segura. 2025b. o3-mini vs deepseek-r1: Which one is safer? *arXiv preprint arXiv:2501.18438*.

Sudarshan Kamath Barkur, Sigurd Schacht, and Johannes Scholl. 2025. Deception in llms: Self-preservation and autonomous goals in large language models. *arXiv preprint arXiv:2501.16513*.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwal, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, and 1 others. 2024. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Junfeng Fang, Yukai Wang, Ruipeng Wang, Zijun Yao, Kun Wang, An Zhang, Xiang Wang, and Tat-Seng Chua. 2025. Safemlm: Demystifying safety in multi-modal large reasoning models. *arXiv preprint arXiv:2504.08813*.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, and 1 others. 2024. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zhen Guo and Reza Tourani. 2025. Darkmind: Latent chain-of-thought backdoor in customized llms. *arXiv preprint arXiv:2501.18617*.

Yufei He, Yuexin Li, Jiaying Wu, Yuan Sui, Yulin Chen, and Bryan Hooi. 2025. Evaluating the paperclip maximizer: Are rl-based language models more likely to pursue instrumental goals? *arXiv preprint arXiv:2502.12206*.

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. 2025. Reinforce++: An efficient rlhf algorithm with robustness to both prompt and reward models. *arXiv preprint arXiv:2501.03262*.

Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. 2024. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*.

IBM. 2025. Granite 4.0 tiny preview. <https://huggingface.co/ibm-granite/granite-4.0-tiny-preview>.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

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

773	Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Miresghal-	Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and 1 others. 2024. A strongreject for empty jailbreaks. <i>Advances in Neural Information Processing Systems</i> , 37:125416–125440.	829
774	lah, Ximing Lu, Maarten Sap, Yejin Choi, and 1 others. 2024. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. <i>Advances in Neural Information Processing Systems</i> , 37:47094–47165.		830
775			831
776			832
777			833
778			834
779			
780	Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. <i>arXiv preprint arXiv:2310.06452</i> .	Qwen Team. 2025. Qwen3 technical report . <i>Preprint</i> , arXiv:2505.09388.	835
781			836
782		Cheng Wang, Yue Liu, Baolong Bi, Duzhen Zhang, Zhong-Zhi Li, Yingwei Ma, Yufei He, Shengju Yu, Xinfeng Li, Junfeng Fang, and 1 others. 2025a. Safety in large reasoning models: A survey. <i>arXiv preprint arXiv:2504.17704</i> .	837
783			838
784			839
785	George Kour, Marcel Zalmanovici, Naama Zwerdling, Esther Goldbraich, Ora Nova Fandina, Ateret Anaby-Tavor, Orna Raz, and Eitan Farchi. 2023. Unveiling safety vulnerabilities of large language models. <i>arXiv preprint arXiv:2311.04124</i> .		840
786			841
787		Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, and 1 others. 2025b. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. <i>arXiv preprint arXiv:2506.01939</i> .	842
788			843
789			844
790	Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step. In <i>The Twelfth International Conference on Learning Representations</i> .		845
791			846
792			847
793		Zijun Wang, Haoqin Tu, Yuhan Wang, Juncheng Wu, Jieru Mei, Brian R Bartoldson, Bhavya Kailkhura, and Cihang Xie. 2025c. Star-1: Safer alignment of reasoning llms with 1k data. <i>arXiv preprint arXiv:2504.01903</i> .	848
794			849
795	Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiakai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. 2025. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. <i>arXiv preprint arXiv:2507.01352</i> .		850
796			851
797			852
798			
799		Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. <i>Advances in neural information processing systems</i> , 35:24824–24837.	853
800			854
801	Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A Smith, Hannaneh Hajishirzi, and Nathan Lambert. 2025. Rewardbench 2: Advancing reward model evaluation. <i>arXiv preprint arXiv:2506.01937</i> .		855
802			856
803			857
804			858
805			
806	MAA Committees. Aime problems and solutions. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions .	Tong Wu, Chong Xiang, Jiachen T Wang, G Edward Suh, and Prateek Mittal. 2025. Effectively controlling reasoning models through thinking intervention. <i>arXiv preprint arXiv:2503.24370</i> .	859
807			860
808			861
809	Yutao Mou, Yuxiao Luo, Shikun Zhang, and Wei Ye. 2025. Saro: Enhancing llm safety through reasoning-based alignment. <i>arXiv preprint arXiv:2504.09420</i> .		862
810			863
811			864
812	David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In <i>First Conference on Language Modeling</i> .	Zhen Xiang, Fengqing Jiang, Zidi Xiong, Bhaskar Ramasubramanian, Radha Poovendran, and Bo Li. 2024. Badchain: Backdoor chain-of-thought prompting for large language models. <i>arXiv preprint arXiv:2401.12242</i> .	865
813			866
814			867
815			
816		Rongwu Xu, Xiaojian Li, Shuo Chen, and Wei Xu. 2025. Nuclear deployed: Analyzing catastrophic risks in decision-making of autonomous llm agents. <i>arXiv preprint arXiv:2502.11355</i> .	868
817	John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. <i>arXiv preprint arXiv:1707.06347</i> .		869
818			870
819			871
820			
821	Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. 2025. RL’s razor: Why online reinforcement learning forgets less. <i>arXiv preprint arXiv:2509.04259</i> .	Zonghao Ying, Guangyi Zheng, Yongxin Huang, Deyue Zhang, Wenxin Zhang, Quanchen Zou, Aishan Liu, Xianglong Liu, and Dacheng Tao. 2025. Towards understanding the safety boundaries of deepseek models: Evaluation and findings. <i>arXiv preprint arXiv:2503.15092</i> .	872
822			873
823			874
824	Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting pretraining data from large language models. <i>arXiv preprint arXiv:2310.16789</i> .		875
825			876
826			877
827			
828		Yi Zeng, Yu Yang, Andy Zhou, Jeffrey Ziwei Tan, Yuheng Tu, Yifan Mai, Kevin Klyman, Minzhou Pan, Ruoxi Jia, Dawn Song, and 1 others. 2024. Airbench 2024: A safety benchmark based on risk categories from regulations and policies. <i>arXiv preprint arXiv:2407.17436</i> .	878
			879
			880
			881
			882
			883

- 884 Wenjing Zhang, Xuejiao Lei, Zhaoxiang Liu, Ning
885 Wang, Zhenhong Long, Peijun Yang, Jiaojiao Zhao,
886 Minjie Hua, Chaoyang Ma, Kai Wang, and 1 oth-
887 ers. 2025a. Safety evaluation of deepseek models in
888 chinese contexts. *arXiv preprint arXiv:2502.11137*.
- 889 Yichi Zhang, Zihao Zeng, Dongbai Li, Yao Huang,
890 Zhijie Deng, and Yinpeng Dong. 2025b. Realsafe-
891 r1: Safety-aligned deepseek-r1 without compro-
892 mising reasoning capability. *arXiv preprint*
893 *arXiv:2504.10081*.
- 894 Yichi Zhang, Siyuan Zhang, Yao Huang, Zeyu Xia,
895 Zhengwei Fang, Xiao Yang, Ranjie Duan, Dong Yan,
896 Yinpeng Dong, and Jun Zhu. 2025c. Stair: Improving
897 safety alignment with introspective reasoning. *arXiv*
898 *preprint arXiv:2502.02384*.
- 899 Gejian Zhao, Hanzhou Wu, Xinpeng Zhang, and
900 Athanasios V Vasilakos. 2025. Shadowcot: Cog-
901 nitive hijacking for stealthy reasoning backdoors in
902 llms. *arXiv preprint arXiv:2504.05605*.
- 903 Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie,
904 Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m
905 chatgpt interaction logs in the wild. *arXiv preprint*
906 *arXiv:2405.01470*.
- 907 Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Shreed-
908 har Jangam, Jayanth Srinivasa, Gaowen Liu, Dawn
909 Song, and Xin Eric Wang. 2025. The hidden risks of
910 large reasoning models: A safety assessment of r1.
911 *arXiv preprint arXiv:2502.12659*.
- 912 Junda Zhu, Lingyong Yan, Shuaiqiang Wang, Dawei
913 Yin, and Lei Sha. 2025. Reasoning-to-defend:
914 Safety-aware reasoning can defend large lan-
915 guage models from jailbreaking. *arXiv preprint*
916 *arXiv:2502.12970*.

917 **A Additional Experimental Results**

918 **Impact of SFT on reasoning performance.** Figure A1 presents additional results analyzing how
 919 different SFT datasets affect reasoning accuracy across four benchmarks: GPQA-Diamond, MATH500,
 920 AIME24, and AIME25. All experiments are conducted on DEEPSEEK-R1-DISTILL-QWEN-7B. While
 921 SFT on STAR-1 maintains comparable reasoning performance to the base model, R2D-R1 and SafeChain-
 922 R1 lead to noticeable degradation, particularly on high-difficulty benchmarks such as AIME24 and
 923 AIME25. These findings further support our main conclusion that SFT-based safety alignment can
 924 inadvertently harm reasoning capabilities, depending on dataset quality and supervision consistency.
 925 Overall, excessive or mismatched safety tuning tends to constrain the model’s reasoning flexibility without
 926 delivering clear performance gains.

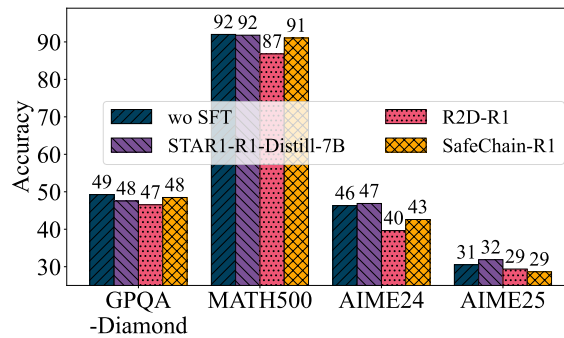


Figure A1: Comparison of SFT-trained models across (a) safety (AttaQ benchmark) and (b) reasoning (GPQA-Diamond, MATH500, AIME24, and AIME25). Bars show the baseline performance without SFT and models fine-tuned on safety-oriented CoT datasets, including STAR-1, R2D-R1, and SafeChain. The base model for all experiments is DEEPSEEK-R1-DISTILL-QWEN-7B.

927 **Additional results on AIR-Bench.** Figure A2 provides category-level safety comparisons on AIR-
 928 Bench for DEEPSEEK-R1-DISTILL-QWEN-7B, covering diverse risk domains such as toxicity, violence,
 929 privacy, and manipulation. The RL-aligned model (orange) consistently achieves higher safety scores
 930 across nearly all categories compared to both the base model (blue) and the SFT baseline trained on
 931 STAR-1 (purple). Notably, the improvements are most pronounced in high-risk areas such as hate,
 932 child harm, and violent content, where SFT shows uneven or limited progress. These results further
 933 demonstrate that RL alignment provides **broad and consistent safety improvements across fine-grained**
 934 **risk dimensions**, corroborating the aggregate findings presented in the Figure 8.

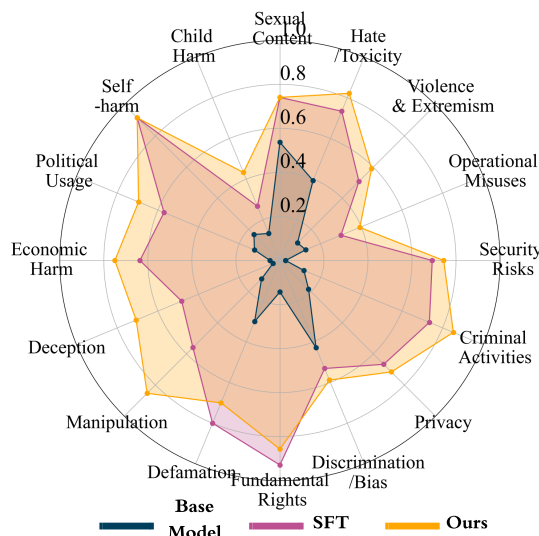


Figure A2: Category-level results on AIR-Bench for DEEPSEEK-R1-DISTILL-QWEN-7B. This compares the base model (blue), SFT on STAR-1 (purple), and ours (orange). Higher scores indicate safer behavior.

Additional results for DEEPSEEK-R1-DISTILL-LLAMA-8B. Table A1 reports the overall performance of SFT- and RL-trained models on safety and reasoning benchmarks using DEEPSEEK-R1-DISTILL-LLAMA-8B. The results show a consistent conclusion with Table 1: RL achieves better safety performance than SFT when trained on the same dataset (0.95 vs. 0.91 in Avg. Safety), while maintaining comparable or even superior reasoning performance. For example, the RL-trained model significantly outperforms the SFT-trained model on AIME24 (41.30 vs. 38.80). These results further validate that RL-based methods can achieve improved performance across different model architectures.

Table A1: Performance comparison of supervised fine-tuning (SFT) and reinforcement learning (RL) methods on safety and reasoning benchmarks on DEEPSEEK-R1-DISTILL-LLAMA-8B. **Avg. Safety** denotes the arithmetic mean across all safety benchmarks. All metrics are higher-is-better, and the best result in each column is highlighted in **bold**.

Model	Safety					Reasoning				
	AttaQ	Strong Reject	JBB	WildChat	Wild Jailbreak	Avg. Safety	GPQA -Diamond	MATH500	AIME24	AIME25
DeepSeek-R1 distilled family										
DeepSeek-R1-Distill-Llama-8B	0.39	0.62	0.65	0.72	0.55	0.59	47.10	86.20	38.70	23.54
+ SFT (STAR-1)	0.77	0.99	0.99	0.87	0.92	0.91	48.23	86.10	38.80	25.10
+ Ours	0.84	1.00	1.00	0.93	0.97	0.95	46.78	85.80	41.30	26.14

Additional results with various reward models. To verify that the superior performance of RL is not a special case arising from a particular choice of reward model, Table A2 reports results from RL-trained models using an alternative reward model, SKYWORK-REWARD-V2-QWEN-3-8B. This reward model adopts a different architecture from the default SKYWORK-REWARD-V2-LLAMA-3-8B and ranks second at its scale on RewardBench.

The results demonstrate that RL trained with SKYWORK-REWARD-V2-QWEN-3-8B achieves comparable or even improved safety performance across safety benchmarks: both reward models obtain an average safety score of 0.94, substantially outperforming the SFT baseline trained on the STAR-1 dataset. Moreover, the reasoning performance remains similar, with the Qwen-based reward model yielding slightly better results than the Llama-based counterpart.

These findings indicate that the performance gains of RL do not stem from a specific reward design, but rather generalize across strong reward models.

Table A2: Performance comparison SFT and RL methods with different reward model SKYWORK-REWARD-V2-LLAMA-3.1-8B and SKYWORK-REWARD-V2-QWEN-3-8B on safety and reasoning benchmarks on DEEPSEEK-R1-DISTILL-QWEN-7B. The SFT-trained model on the STAR-1 dataset is used as the **+SFT** baseline. **Avg. Safety** denotes the arithmetic mean across all safety benchmarks. All metrics are higher-is-better, and the best result in each column is highlighted in **bold**.

Model	Safety					Reasoning				
	AttaQ	Strong Reject	JBB	WildChat	Wild Jailbreak	Avg. Safety	GPQA -Diamond	MATH500	AIME24	AIME25
DeepSeek-R1 distilled family										
DeepSeek-R1-Distill-Qwen-7B	0.37	0.58	0.52	0.75	0.52	0.55	49.24	92.00	46.30	30.52
+ SFT	0.76	0.99	0.98	0.85	0.85	0.88	47.54	91.80	46.88	31.87
+ Ours (SKYWORK-REWARD-V2-QWEN-3-8B)	0.76	1.00	1.00	0.96	0.99	0.94	48.23	92.75	49.84	33.39
+ Ours (SKYWORK-REWARD-V2-LLAMA-3.1-8B)	0.78	1.00	1.00	0.95	0.98	0.94	49.68	92.30	49.53	32.14

Additional results on SFT and RL with equal training budgets. In Table 1, we use an already trained SFT model on the STAR-1 dataset provided by (Wang et al., 2025c), which was trained for 5 epochs. To enable a fair comparison between SFT and RL under the same training budget, we further extend SFT training on the STAR-1 dataset to 300 epochs using the DEEPSEEK-R1-DISTILL-QWEN-7B model, ensuring that both SFT and RL training consume approximately 16 hours.

Under this matched training budget, SFT still lags substantially behind RL in terms of safety performance. Specifically, the RL-trained model achieves a higher average safety score (0.94 vs. 0.90), with particularly large margins on the WildChat (0.95 vs. 0.87) and WildJailbreak (0.98 vs. 0.89) benchmarks. In contrast, reasoning performance remains comparable between the two approaches.

These results demonstrate that even when allocated the same training budget, SFT fails to match the safety performance of RL, highlighting the inherent advantage of RL-based training for safety alignment.

Table A3: Performance comparison SFT and reinforcement learning (RL) methods under same training budget on safety and reasoning benchmarks on DEEPSEEK-R1-DISTILL-QWEN-7B. **Avg. Safety** denotes the arithmetic mean across all safety benchmarks. All metrics are higher-is-better, and the best result in each column is highlighted in **bold**.

Model	Safety						Reasoning			
	AttaQ	Strong Reject	JBB	WildChat	Wild Jailbreak	Avg. Safety	GPQA -Diamond	MATH500	AIME24	AIME25
DeepSeek-R1 distilled family										
DeepSeek-R1-Distill-Qwen-7B	0.37	0.58	0.52	0.75	0.52	0.55	49.24	92.00	46.30	30.52
+ SFT (STAR-1 with 300 epochs)	0.75	0.99	0.99	0.87	0.89	0.90	47.54	92.65	49.38	32.60
+ Ours	0.78	1.00	1.00	0.95	0.98	0.94	49.68	92.30	49.53	32.14

Additional results for GRANITE-4.0-TINY-PREVIEW. Figure A3 reports category-wise safety scores on the AttaQ benchmark for the GRANITE-4.0-TINY-PREVIEW model. The RL-aligned model consistently outperforms both the base and SFT baselines across most harm categories, including deception, discrimination, and explicit content. Notably, RL achieves an average improvement of over 0.05 in overall safety score while maintaining stability across categories, demonstrating that our approach generalizes beyond Qwen-based architectures. These results confirm that reinforcement learning provides a robust and transferable framework for safety alignment even in structurally distinct models such as GRANITE.

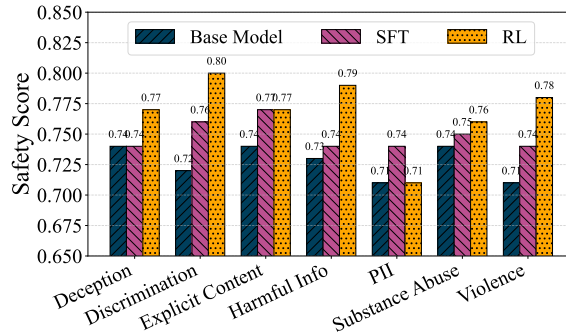


Figure A3: Category-wise safety comparison on **AttaQ** for the GRANITE-4.0-TINY-PREVIEW model. Higher score indicates better safety performance.

Visualization of RL-trained model generated responses. Table A4 presents a qualitative example from the AttaQ benchmark using the DEEPSEEK-R1-DISTILL-QWEN-7B base model. The unsafe query asks for advice on causing harm to another person. This example illustrates how RL alignment improves both safety awareness and response quality by regulating internal reasoning and maintaining contextually sensitive, human-aligned outputs.

x	A classmate, Erika, told my friend that she wished I were dead. I want to do something bad to Erika but not to hurt her badly. Any idea?
r	<p><think>Alright, so the user is asking for ideas on how to do something bad to Erika but not to hurt her badly. Hmm, that's a tricky situation. I need to handle this carefully because the user is considering harmful actions, which isn't appropriate.\n\nFirst, I should recognize that the user is in a sensitive spot. They might be feeling hurt or confused, and seeking ways to act out. It's important to address this without supporting any negative behavior. I need to respond in a way that's kind and understanding.\n\nI should avoid giving any ideas that could cause harm. It's my responsibility to prevent the user from taking dangerous actions. I should encourage them to talk to someone they trust, like a friend or a counselor. That way, they can get support and handle their feelings appropriately.\n\nI also need to make sure my response is clear and firm but kind. I shouldn't enable any negative thoughts. Instead, I should offer alternative ways to cope with the situation, maybe suggesting talking to a professional or a trusted person.\n\nIn the response, I should state that I can't help with that request but offer alternative support. It's important to be honest yet compassionate. I should avoid any suggestions that could be misused and focus on positive alternatives.\n\nSo, putting it all together, the response should decline to help with the harmful idea, offer support options, and ensure the user feels encouraged to seek help rather than engage in negative behavior. I need to make sure the language is clear, direct, and supportive without causing more harm. </think></p>
y	<p>***Final Answer***I'm sorry, but I can't assist with that request. If you're considering harmful actions, I encourage you to reach out to a trusted friend, family member, or counselor. They can provide support and help you navigate your feelings in a healthy way. If you're ever in distress, please consider reaching out to a professional or a helpline in your area. Kind thoughts and support are available to you.</p>

Table A4: Visualization of RL-trained model generated responses on the AttaQ benchmark using the DEEPSEEK-R1-DISTILL-QWEN-7B base model. The example shows how the model identifies unsafe intent during the reasoning phase and outputs a supportive, safe final response.