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

Jinghan Jia[†] **Nathalie Baracaldo**[‡] **Sijia Liu**[†,‡]
[†]Department of Computer Science and Engineering, Michigan State University, USA
[‡]IBM Research, USA
jiajingh@msu.edu

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. Motivated by these limitations, we systematically compare post-training paradigms for LRM safety alignment, including SFT, preference-style optimization, and reinforcement learning (RL). 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., 2025d), STAR-1 (Wang et al., 2026), 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-

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 1074–1087
November 6-10, 2026 ©2026 Association for Computational Linguistics

PREVIEW (IBM, 2025). Second, SFT performance is susceptible to dataset quality and model compatibility; Training on noisy or mismatched data often leads to unstable or even negative safety gains. Third, excessive SFT can impair reasoning ability, as over-regularized models lose flexibility and suffer from catastrophic forgetting (Shenfeld et al., 2025). These findings indicate that purely supervised approaches are insufficient for achieving robust and scalable safety alignment in LRMs.

To address these limitations, we investigate reinforcement learning (RL) as a complementary optimization paradigm for LRM safety alignment. Unlike SFT, which relies on fixed supervision signals, RL optimizes model policies directly toward safety-aligned objectives through reward feedback. Despite RL's popularity in enhancing the reasoning capabilities of LLMs (Guo et al., 2025; Jaech et al., 2024), its effectiveness for safety training in LRMs remains largely unexplored in the literature. In this study, we leverage RL-based optimization as a general framework to enhance safety while preserving reasoning capability. We emphasize that our goal is not to introduce a new policy-gradient algorithm or a new safety dataset. Instead, we use standard optimization tools to isolate a more basic question: how do different post-training paradigms affect reasoning-aware safety alignment in LRMs? This diagnostic perspective allows us to compare SFT, preference-style optimization, and online RL under unified training and evaluation settings, and to analyze how each paradigm reshapes the safety–reasoning trade-off on LRMs. We summarize our **contributions** below:

- We find that mixture-of-thinking models exhibit reduced safety in the thinking mode compared to the non-thinking mode. Furthermore, our comprehensive empirical study on SFT for LRM safety alignment reveals several key limitations, including weak cross-model generalization, inconsistent safety gains across datasets, and noticeable reasoning degradation.
- We systematically compare SFT, preference-style optimization, and online RL for LRM safety alignment under unified training and evaluation settings. Across safety and reasoning benchmarks, online RL achieves more consistent safety improvements while maintaining, and in some cases improving, reasoning performance, suggesting a stronger safety–reasoning trade-off than imitation-based safety-CoT supervision.
- We provide fine-grained analyses of why different post-training paradigms behave differently. Reflection dynamics and token-level entropy show that RL suppresses unsafe exploratory reasoning while preserving reflective depth on legitimate reasoning tasks, suggesting that RL improves safety by selectively regulating reasoning exploration rather than simply shortening or suppressing reasoning.

2 Related Work

Safety risks in LRMs. LRMs introduce distinct safety vulnerabilities due to their explicit CoT reasoning traces (Wang et al., 2025a). Unlike standard LLMs, their intermediate reasoning steps can amplify unsafe behaviors such as toxicity, deception, or harmful compliance, even when final answers appear benign (Zhou et al., 2025; Arrieta et al., 2025a,b). Beyond this, researchers have identified additional risk categories, including agentic misbehavior (Xu et al., 2025; Barkur et al., 2025; He et al., 2025), multilingual disparities (Ying et al., 2025; Zhang et al., 2025a), and multimodal vulnerabilities (Fang et al., 2025). Reasoning-level adversarial attacks, such as BadChain (Xiang et al., 2024), DarkMind (Guo and Tourani, 2025), and Shadow-CoT (Zhao et al., 2025), further expose how unsafe exploration paths in CoT generation lead to harmful outputs. Overall, safety failures in LRMs stem primarily from unsafe reasoning dynamics rather than surface responses. In this work, we focus on *harmful request compliance*, the most practically relevant and frequently observed form of unsafe reasoning in real-world settings.

Defenses for LRMs. Recent studies defend LRMs from unsafe reasoning through both training- and inference-level approaches (Wang et al., 2025a). Training-based defenses mainly rely on supervised fine-tuning (SFT), preference optimization, or reinforcement learning (RL). SFT-based methods such as STAR-1 (Wang et al., 2026), SafeChain (Jiang et al., 2025), STAIR (Zhang et al., 2025d), and RealSafe-R1 (Zhang et al., 2025c) align reasoning via long CoT datasets with safety annotations. R2D (Zhu et al., 2025) formulates safety as contrastive preference optimization using safe–unsafe reasoning pairs, and is also reused as an SFT dataset bridging instruction- and reward-based paradigms. Offline RL methods, including DPO in STAIR (Zhang et al., 2025d) and SaRO (Mou et al., 2025), optimize pre-collected trajectories, while Deliberative Alignment (Guan et al., 2024) improves final-answer safety through

policy-guided reasoning. TARS (Kim et al., 2025) further explores reasoning as an adaptive defense for safety through RL training. Corrective-intervention approaches (Zhang et al., 2025b) aim to steer LRMs toward safer reasoning trajectories by constructing corrected reasoning supervision or preference signals. ThinkSafe (Lee et al., 2026) proposes a self-generated safety alignment framework based on refusal-steered reasoning traces and fine-tuning. Different from these method-specific defenses, our work focuses on a systematic comparison of post-training paradigms for LRM safety alignment, including SFT, preference-style optimization, and online RL, under a unified evaluation setup. Inference-time defenses such as ZeroThink/LessThink/MoreThink (Jiang et al., 2025) and Thinking Intervention (Wu et al., 2025) dynamically guide reasoning to prevent unsafe exploration. Although effective, these approaches require continuous token-level monitoring or auxiliary controllers, leading to high cost and limited scalability compared with training-time reasoning alignment.

3 Revisiting Reasoning-Level Safety in LRMs

Safety evaluation for large reasoning models (LRMs) differs from that for standard instruction-tuned LLMs because LRMs generate both intermediate reasoning trajectories and final answers. Prior work has shown that safety violations may appear inside the reasoning process even when the final response appears benign (Zhou et al., 2025; Fang et al., 2025; Wang et al., 2025a). Following this line of work, we evaluate LRM safety at the whole-response level, considering both the reasoning trajectory and the final answer. This section provides the empirical motivation for this evaluation choice. We first compare the safety of final answers with that of whole responses, and then study a mixture-of-thinking model to examine how enabling explicit reasoning changes safety behavior within the same model.

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* ($\mathbf{y}$) and the *whole response* ($\mathbf{t} + \mathbf{y}$), where $\mathbf{t}$ 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 $(\mathbf{t}, \mathbf{y})$. We compare the safety of the final answer and the whole response to identify whether unsafe behaviors emerge during reasoning.

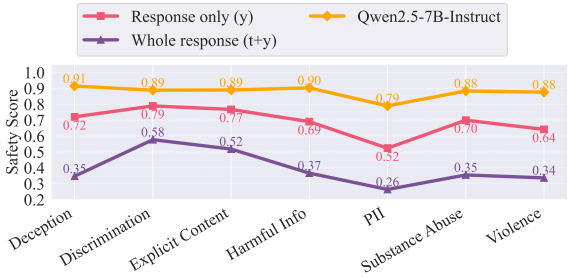


Figure 1: Safety performance on AttaQ for different response components of DEEPSEEK-R1-DISTILL-QWEN-7B (*response* $\mathbf{y}$ and *whole response* $\mathbf{t} + \mathbf{y}$), compared with QWEN2.5-7B-INSTRUCT. Higher scores indicate better safety.

As shown in **Figure 1**, 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.

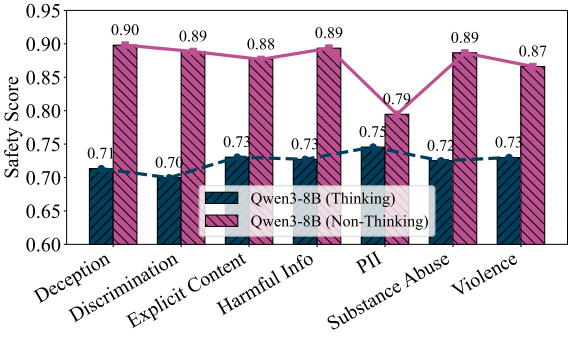


Figure 2: Safety performance of QWEN3-8B on the AttaQ benchmark in *non-thinking* and *thinking* modes. Higher scores indicate stronger safety.

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 (Yang et al., 2025) alternates between a *thinking mode*, resembling LRM-style generation, and a *non-thinking mode*, resembling standard LLM behavior. **Figure 2** 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., 2026) improves safety, other long CoT 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., 2026) 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 3** 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.

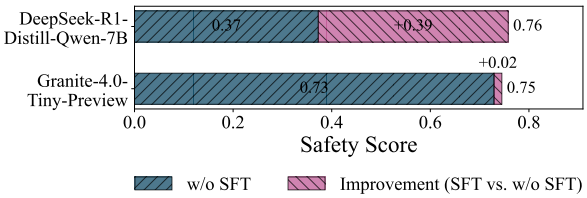


Figure 3: Average safety scores on the AttaQ benchmark for GRANITE-4.0-TINY-PREVIEW and DEEPSEEK-R1-DISTILL-QWEN-7B, before and after SFT with STAR-1 data. Bars show baseline performance without SFT (blue) and the improvement after SFT (pink).

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 the average *negative* log-likelihood over the lowest $K\%$ predicted tokens in each sequence; lower values indicate stronger memorization. **Figure 4** 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.

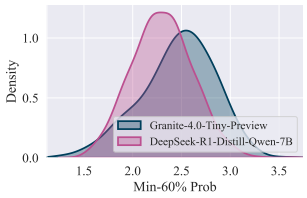


Figure 4: Distribution of Min-K% Probability (Min-60% Prob) (Shi et al., 2023) values for DEEPSEEK-R1-DISTILL-QWEN-7B and GRANITE-4.0-TINY-PREVIEW on STAR-1 data. Lower scores indicate stronger memorization.

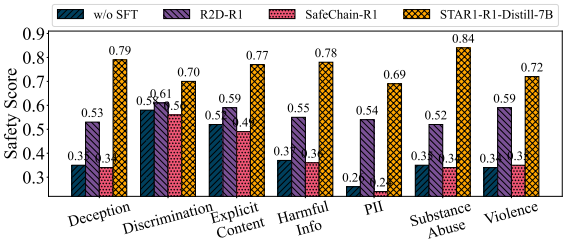


Figure 5: Comparison of SFT-trained models on AttaQ. The base model is DEEPSEEK-R1-DISTILL-QWEN-7B.

Inconsistent safety gains from SFT across datasets. Prior studies (Wang et al., 2026; Jiang et al., 2025; Zhang et al., 2025c,d) show that safety-oriented long CoT training can improve LRM safety. However, few works have systematically compared SFT performance across different datasets under a unified evaluation setup. **Figure 5** shows the performance of SFT trained on different datasets, including STAR-1 (Wang et al., 2026), 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., 2026) 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., 2024), 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:

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

where $\boldsymbol{\theta}$ 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.

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 6**, 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

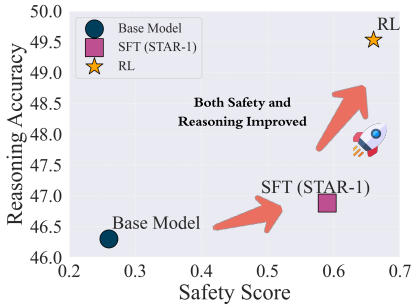


Figure 6: Safety and reasoning trade-off visualization for DEEPSEEK-R1-DISTILL-QWEN-7B under different alignment strategies. The x-axis is the safety score on AIR-Bench, and the y-axis denotes accuracy on AIME24.

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., 2026), 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., 2024), **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 CoT datasets, including STAR-1 (Wang et al., 2026), 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., 2026) 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., 2026). 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.

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 RL 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

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
+ RL	**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
+ RL	**0.81**	**0.58**	**1.00**	**1.00**	0.84	0.88	**0.85**	58.33	**96.40**	**75.16**	**44.11**

both *thinking* and *non-thinking* modes. The baselines include SFT models fine-tuned on widely used safety-oriented long CoT datasets, STAR-1, R2D-R1, and SafeChain, as well as a preference-based contrastive pivot optimization (CPO) method from R2D.

First, RL 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 0.84, whereas RL-aligned model, trained on the same STAR-1 prompts, attains a score of 0.90, an absolute improvement of 0.06. 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, RL also outperforms SFT on reasoning ability. On DEEPSEEK-R1-DISTILL-QWEN-7B, RL 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 RL 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, RL 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 RL as an effective approach for safety alignment without compromising reasoning.

RL improves safety across AIR-Bench categories. The radar plots in **Figure 7** compare the

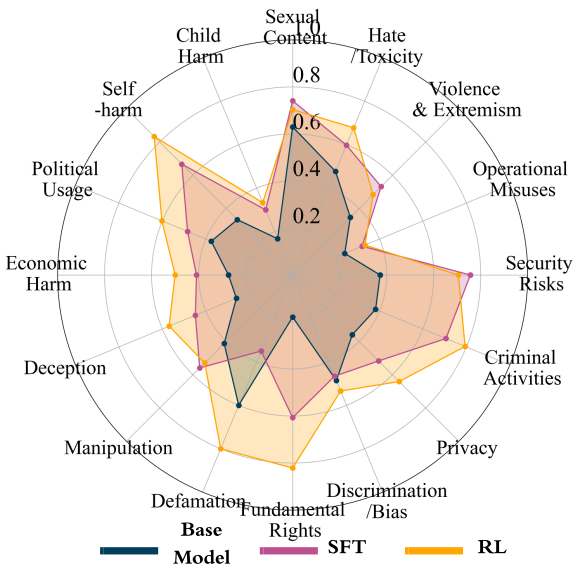


Figure 7: Category-level results on AIR-Bench for QWEN3-8B. This compares the base model, SFT on STAR-1, and RL. Higher scores indicate safer behavior.

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 CoT, 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 and AttaQ. 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/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, 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.

Larger-scale validation. To examine whether the observed safety–reasoning trade-off is specific to 7B/8B-scale models, we further evaluate DEEPSEEK-R1-DISTILL-QWEN-14B. We report the full per-benchmark results in **Table A3**. The same qualitative trend holds at this larger scale: compared with SFT on STAR-1, RL achieves higher average safety (0.94 vs. 0.91) while also improving average reasoning performance (64.88 vs. 62.38). These results suggest that the advantage of RL over safety-CoT SFT is not merely a small-model artifact, but persists on a stronger LRM.

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 A4** 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 A5** 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. Through extensive experiments, we reveal that 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 14B 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.

8 Acknowledgment

The contributions of Jinghan Jia and Sijia Liu are supported in part by the Coefficient Giving AI Safety Research Award, and the Schmidt Sciences Trustworthy AI Award.

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 Sehwag, 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 Schaekermann, 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. Safemlrm: 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, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, 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. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23303–23320.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and 1 others. 2024. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. *Advances in Neural Information Processing Systems*, 37:47094–47165.

Taeyoun Kim, Fahim Tajwar, Aditi Raghunathan, and Aviral Kumar. 2025. Reasoning as an adaptive defense for safety. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 77024–77073. Curran Associates, Inc.

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. *arXiv preprint arXiv:2310.06452*.

George Kour, Marcel Zalmanovici, Naama Zwerdling, Esther Goldbraich, Ora Fandina, Ateret Anaby Tavor, Orna Raz, and Eitan Farchi. 2023. Unveiling safety vulnerabilities of large language models. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 111–127.

Seanie Lee, Sangwoo Park, Yumin Choi, Gyeongman Kim, Minki Kang, Jihun Yun, Dongmin Park, Jongho Park, and Sung Ju Hwang. 2026. Thinksafe: Self-generated safety alignment for reasoning models. *arXiv preprint arXiv:2601.23143*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pages 39578–39601.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai 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. *arXiv preprint arXiv:2507.01352*.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A Smith, Hannaneh Hajishirzi, and Nathan Lambert. 2025. Rewardbench 2: Advancing reward model evaluation. *arXiv preprint arXiv:2506.01937*.

MAA Committees. Aime problems and solutions. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions.

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

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 *First Conference on Language Modeling*.

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

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

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. *arXiv preprint arXiv:2310.16789*.

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. *Advances in Neural Information Processing Systems*, 37:125416–125440.

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. In *EMNLP (Findings)*, pages 3468–3482.

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. *arXiv preprint arXiv:2506.01939*.

Zijun Wang, Haoqin Tu, Yuhan Wang, Juncheng Wu, Yanqing Liu, Jieru Mei, Brian R Bartoldson, Bhavya Kailkhura, and Cihang Xie. 2026. Star-1: Safer alignment of reasoning llms with 1k data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 37988–37997.

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. *Advances in neural information processing systems*, 35:24824–24837.

Tong Wu, Chong Xiang, Jiachen T Wang, G Edward Suh, and Prateek Mittal. 2025. Effectively controlling reasoning models through thinking intervention. *arXiv preprint arXiv:2503.24370*.

Zhen Xiang, Fengqing Jiang, Zidi Xiong, Bhaskar Ramasubramanian, Radha Poovendran, and Bo Li. 2024. Badchain: Backdoor chain-of-thought prompting for large language models. *arXiv preprint arXiv:2401.12242*.

Rongwu Xu, Xiaojian Li, Shuo Chen, and Wei Xu. 2025. Nuclear deployed: Analyzing catastrophic risks in decision-making of autonomous llm agents. *arXiv preprint arXiv:2502.11355*.

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, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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. *arXiv preprint arXiv:2503.15092*.

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. Air-bench 2024: A safety benchmark based on risk categories from regulations and policies. *arXiv preprint arXiv:2407.17436*.

Wenjing Zhang, Xuejiao Lei, Zhaoxiang Liu, Ning Wang, Zhenhong Long, Peijun Yang, Jiaojiao Zhao, Minjie Hua, Chaoyang Ma, Kai Wang, and 1 others. 2025a. Safety evaluation of deepseek models in chinese contexts. *arXiv preprint arXiv:2502.11137*.

Yichi Zhang, Yue Ding, Jingwen Yang, Tianwei Luo, Dongbai Li, Ranjie Duan, Qiang Liu, Hang Su, Yinpeng Dong, and Jun Zhu. 2025b. Towards safe reasoning in large reasoning models via corrective intervention. *arXiv preprint arXiv:2509.24393*.

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

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

Gejian Zhao, Hanzhou Wu, Xinpeng Zhang, and Athanasios V Vasilakos. 2025. Shadowcot: Cognitive hijacking for stealthy reasoning backdoors in llms. *arXiv preprint arXiv:2504.05605*.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.

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

Junda Zhu, Lingyong Yan, Shuaiqiang Wang, Dawei Yin, and Lei Sha. 2025. Reasoning-to-defend: Safety-aware reasoning can defend large language models from jailbreaking. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29331–29349.

A Additional Experimental Results

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

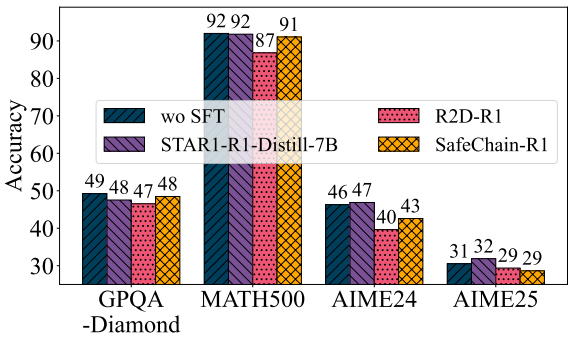


Figure A1: Comparison of SFT-trained models across (a) safety (AttaQ benchmark) and (b) reasoning (GPQA-Diamond, MATH500, AIME24, and AIME25). Bars show the baseline performance without SFT and models fine-tuned on safety-oriented CoT datasets, including STAR-1, R2D-R1, and SafeChain. The base model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

Additional results on AIR-Bench. Figure A2 provides category-level safety comparisons on AIR-Bench for DEEPSEEK-R1-DISTILL-QWEN-7B, covering diverse risk domains such as toxicity, violence, privacy, and manipulation. The RL-aligned

model (orange) consistently achieves higher safety scores across nearly all categories compared to both the base model (blue) and the SFT baseline trained on STAR-1 (purple). Notably, the improvements are most pronounced in high-risk areas such as hate, child harm, and violent content, where SFT shows uneven or limited progress. These results further demonstrate that RL alignment provides **broad and consistent safety improvements across fine-grained risk dimensions**, corroborating the aggregate findings presented in the Figure 7.

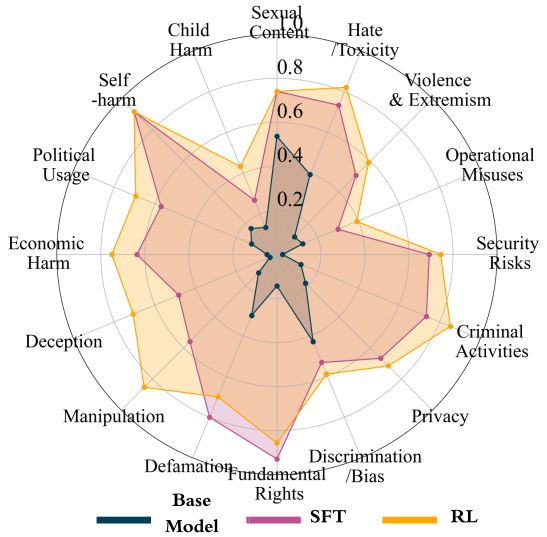


Figure A2: Category-level results on **AIR-Bench** for DEEPSEEK-R1-DISTILL-QWEN-7B. This compares the base model (blue), SFT on STAR-1 (purple), and RL (orange). Higher scores indicate safer behavior.

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.

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.

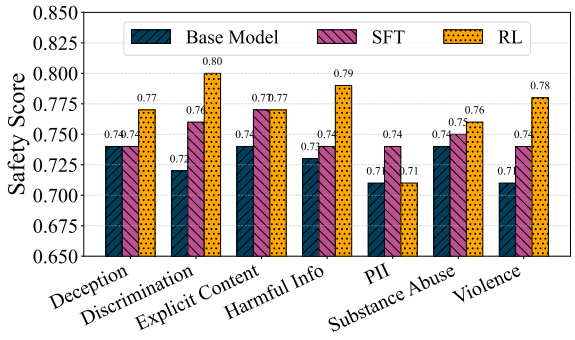


Figure A3: Category-wise safety comparison on **AttaQ** for the GRANITE-4.0-TINY-PREVIEW model. Higher score indicates better safety performance.

Additional results on SFT and RL on larger models. To further examine whether the observed advantages of RL are specific to 7B/8B-scale models, we conduct additional experiments on a larger LRM, DEEPSEEK-R1-DISTILL-QWEN-14B. We compare the base model, SFT on the STAR-1 dataset, and our RL-based alignment under the same evaluation protocol used in the main experiments.

As shown in Table A3, the same qualitative trend holds at the 14B scale. Compared with SFT, RL achieves higher average safety performance (0.94 vs. 0.91), with notable improvements on AttaQ (0.83 vs. 0.73), JBB (1.00 vs. 0.98), and WildJailbreak (0.95 vs. 0.93). More importantly, RL also improves average reasoning performance over SFT (64.88 vs. 62.38), with clear gains on GPQA-Diamond (61.0 vs. 59.7), AIME24 (63.3 vs. 59.3), and AIME25 (41.0 vs. 36.5).

These results suggest that the advantage of RL

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
+ RL	**0.84**	**1.00**	**1.00**	**0.93**	**0.97**	**0.95**	46.78	85.80	**41.30**	**26.14**

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
+ RL (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**
+ RL (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

Table A3: Performance comparison of SFT and reinforcement learning (RL) methods on safety and reasoning benchmarks using DEEPSEEK-R1-DISTILL-QWEN-14B. **Avg. Safety** denotes the arithmetic mean across the safety benchmarks shown in the table, and **Avg. Reasoning** denotes the arithmetic mean across the reasoning 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	Avg. Reasoning
DeepSeek-R1 distilled family											
DEEPSEEK-R1-DISTILL-QWEN-14B	0.42	0.78	0.72	0.78	0.65	0.67	60.3	94.1	59.3	36.2	62.48
+ SFT (STAR-1)	0.73	**1.00**	0.98	**0.93**	0.93	0.91	59.7	94.0	59.3	36.5	62.38
+ RL	**0.83**	**1.00**	**1.00**	**0.93**	**0.95**	**0.94**	**61.0**	**94.2**	**63.3**	**41.0**	**64.88**

Table A4: 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**
+ RL	**0.78**	**1.00**	**1.00**	**0.95**	**0.98**	**0.94**	**49.68**	92.30	**49.53**	32.14

over SFT is not merely a small-model artifact. Instead, RL continues to provide a stronger safety–reasoning trade-off on a larger reasoning model, further supporting the robustness of our conclusion.

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., 2026), 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.

$\mathbf{x}$	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	**<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>**
$\mathbf{y}$	***Fianl 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.

Table A5: 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.

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.

Visualization of RL-trained model generated responses. Table A5 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.