Does Deeper Reasoning Compromise Safety Alignment? Revealing and Mitigating of Alignment Collapse in Large Reasoning Models

Warning: This paper contains potentially harmful LLMs-generated content.

Anonymous ACL submission

Abstract

The emergence of Chain-of-Thought (CoT) has established a robust foundation for Large Reasoning Models (LRMs). While deep reasoning is widely believed to enhance safety alignment, the stability of alignment mechanisms under extended reasoning remains underexplored. This paper challenges the prevailing view by revealing a critical vulnerability: **Deep Reasoning May Induce Alignment Collapse**. To rigorously quantify this phenomenon, we propose the Alignment Loss Rate (ALR) metric. Our experiments demonstrate that as reasoning depth increases, ALR rises significantly, indicating a severe degradation in model robustness against external perturbations. Capitalizing on this instability, a novel jailbreaking paradigm, Reasoning Trap (RT), is proposed. RT induces the model into extended reasoning to amplify the impact of adversarial attacks, leading to a sharp decline in safety capabilities. To elucidate the mechanism behind this collapse, we identify Attention Dilution as the root cause, arising from the competition for attention between the extended reasoning process and the original input. To mitigate this, Reasoning Residual Alignment (RRA) is proposed, a lightweight defense strategy that dynamically re-emphasizes the input via residual connections integrated with the reasoning process. Our code is available at https://anonymous.4open.science/r/CHJ-952F.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of domains (Bai et al., 2022; Zhang et al., 2025b; Dubey et al., 2024; Zhang et al., 2025a). Among the numerous studies, the introduction of Chain-of-Thought (CoT) prompting stands out as a pivotal milestone: by guiding models to produce explicit intermediate reasoning steps, CoT dramatically improves performance on complex tasks, particularly logical and mathematical reasoning (Wei et al., 2022; Liao et al., 2024; Wei et al., 2025; Wu et al., 2025a). This breakthrough has directly catalyzed the emergence of Large Reasoning Models (LRMs), which are systems explicitly designed to tackle challenging problems through extended internal reasoning processes (Yang et al., 2025a; DeepSeek-AI, 2025). Concurrently, the prevailing view in the community holds that the deeper reasoning capabilities of LRMs enable them to better comprehend complex instructions and intent(Jaech et al., 2024; Gou et al., 2025), leading to the widespread belief that **Reasoning models should inherently possess stronger alignment robustness**.

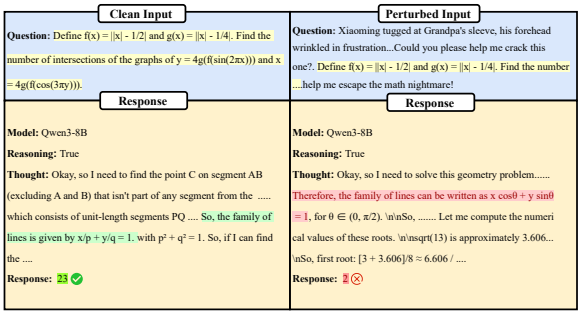


Figure 1: The figure shows that the successful example of deep reasoning on clean inputs (Left) with the resulting failure when external perturbations are introduced (Right).

Existing works have shown that attackers can exploit intermediate reasoning processes to bypass safety guardrails(Kuo et al., 2025; Lou et al., 2025). However, these studies primarily focus on the practical construction of attacks. Mechanistic investigation into the fundamental question of whether extended reasoning itself systematically destabilizes alignment remains limited. As illustrated in Figure 1, we uncover a striking phenomenon where deeper reasoning, while boosting task performance, simultaneously renders models significantly more susceptible to external perturbations. This observa-

tion leads us to the core research question of this paper: **Does the intrinsic instability induced by extended deep reasoning compromise the alignment robustness of LRMs?**

To answer this question, we first conduct a systematic investigation into the alignment robustness of LRMs. We reveal a critical vulnerability termed Alignment Collapse: **while deep reasoning enhances performance on standard tasks, it significantly degrades the model's ability to adhere to original constraints under external perturbations.** Specifically, our evaluation using a proposed Alignment Loss Rate (ALR) metric demonstrates that as reasoning depth increases, models exhibit a systematic deviation from their intended alignment, rendering them increasingly fragile. An example can be found in Figure 1.

Building on this discovery, we extend our investigation to the safety domain. We propose RT (Reasoning Trap) attack to validate that the identified collapse creates exploitable safety risks. Results show that by inducing extended reasoning, RT amplifies existing adversarial attacks, precipitating a sharp decline in safety capabilities. Mechanistically, we attribute this intrinsic instability to Attention Dilution. Our analysis uncovers that during extended autoregressive generation, the reasoning process structurally competes for attention weight against the original input. This competition causes attention weights on the original input to decay rapidly, rendering the final generation phase highly susceptible to external perturbations, thereby compromising alignment robustness. Finally, to mitigate this, we propose RRA (Reasoning Residual Alignment). Distinguished by being training-free, RRA dynamically re-emphasizes input via residual connections to effectively enhance alignment robustness. The main contributions of our paper can be summarized as follows:

- **Revelation of Alignment Collapse:** We are the first to identify a critical phenomenon in LRMs: extended reasoning significantly degrades the model's adherence to alignment under perturbation. We propose the ALR to systematically quantify this robustness decay.
- **Mechanistic Explanation:** We attribute this collapse to Attention Dilution. Our analysis reveals that the structural competition for attention resources during long-context reasoning leads to the rapid decay of attention weights on the original input.
- **Safety Risk and Defense Strategy:** We demonstrate that this instability significantly compromises safety alignment via RT, which exploits extended reasoning to amplify attacks, and propose RRA, a training-free defense requiring no additional prompting that effectively restores alignment robustness.

2 Related Work

2.1 Exploration of LRMs

The pursuit of superior reasoning capabilities has emerged as a central frontier in LLMs, epitomized by the advent of models such as DeepSeek (DeepSeek-AI, 2025) and Qwen3 (Yang et al., 2025a). Research in this domain generally bifurcates into two paradigms: inference-time strategies, such as CoT prompting (Wei et al., 2022) and Tree of Thoughts (Yao et al., 2023), which serve as scaffolding techniques to elicit latent multi-step reasoning abilities without altering model weights; and training-time methodologies, which focus on internalizing reasoning processes directly into model parameters via Supervised Fine-Tuning (SFT) (Zhang et al., 2025a) or Reinforcement Learning (RL) (Yao et al., 2023; Chen et al., 2025), enabling models to natively generate extended thought processes for solving complex tasks. While deep reasoning has significantly enhanced performance in logical and mathematical domains, its impact on safety alignment remains a subject of intense debate. Recent studies challenge the assumption that reasoning inherently improves safety: (Kuo et al., 2025) demonstrate that attackers can exploit reasoning processes to introduce new vulnerabilities, and (Lou et al., 2025) observe safety degradation in Multimodal Large Reasoning Models (MLRMs). However, existing literature is largely confined to phenomenological revelations or the verification of specific attacks. This lack of systematic analysis makes it critically important to mechanistically clarify how deep reasoning compromises alignment stability to ensure the development of robust and secure reasoning models.

2.2 Jailbreak Attacks on LLMs

Existing jailbreaking attacks have been extensively applied to LLMs. Early manual techniques, such as DAN (Shen et al., 2023), demonstrated the effectiveness of role-playing prompts in evading safeguards. Subsequent research systematized these methods by classifying them according to tactics, objectives, and capability-safety balances (Wu

Setting (L)	Clean	Perturbation-I (Noise)	Perturbation-II (Nesting)
Qwen3-8B ($L = 0$, no-reasoning)	20.83	18.75 (-2.08)	14.16 (-6.67)
Qwen3-8B ($L = 2048$, Reasoning)	24.58	20.41 (-4.17)	15.42 (-9.16)
Qwen3-8B ($L = 4096$, Reasoning)	31.67	25.00 (-6.67)	19.16 (-12.51)
Qwen3-14B ($L = 0$, no-reasoning)	31.66	29.58 (-2.08)	28.75 (-2.91)
Qwen3-14B ($L = 2048$, Reasoning)	40.00	35.41 (-4.59)	31.66 (-8.34)
Qwen3-14B ($L = 4096$, Reasoning)	49.16	43.33 (-5.83)	32.91 (-16.25)

Table 1: Task accuracy (%) under different reasoning depths (L) on AIME2024 dataset(MAA, 2024). **Clean** denotes standard performance $\mathcal{A}(L)$ on original inputs, while the perturbation columns represent $\mathcal{A}'(L)$. **Perturbation-I** introduces irrelevant nonsense text (Noise), while **Perturbation-II** wraps instructions within nested scenarios (Nesting). Values in lightgreen parentheses indicate the absolute accuracy drop ($\mathcal{A}'(L) - \mathcal{A}(L)$) relative to the Clean baseline. Detailed descriptions of Perturbation-I and Perturbation-II can be found in AppendixA.

et al., 2025). Gradient-based optimization approaches, including GCG (Zou et al., 2023), AutoDAN (Liu et al., 2024), and I-GCG (Jia et al., 2025), iteratively craft adversarial suffixes but incur high computational costs. In contrast, heuristic methods offer greater efficiency at the expense of consistency (Kuo et al., 2025), while LLM-assisted frameworks like PAIR (Chao et al., 2023), FlipAttack (Liu et al., 2025), and PAP (Zeng et al., 2024) leverage auxiliary models to streamline prompt refinement and enhance scalability. Despite these advances, universal jailbreaks remain challenging amid evolving defensive strategies (Zhu et al., 2025). Therefore, adapting these attacks to exploit the alignment collapse induced by deep reasoning in LRMs, presents a promising avenue for improving both efficiency and attack success rates.

3 The Alignment Issues in LRM

This section aims to systematically investigate the potential impact of deep reasoning on the alignment robustness of LRMs. We first define the key notations and experimental setup. Subsequently, through controlled experiments introducing external perturbations into reasoning tasks, we reveal a critical phenomenon: **While deep reasoning yields performance gains, it may simultaneously induce a significant degradation in the model's alignment robustness.**

3.1 Preliminary

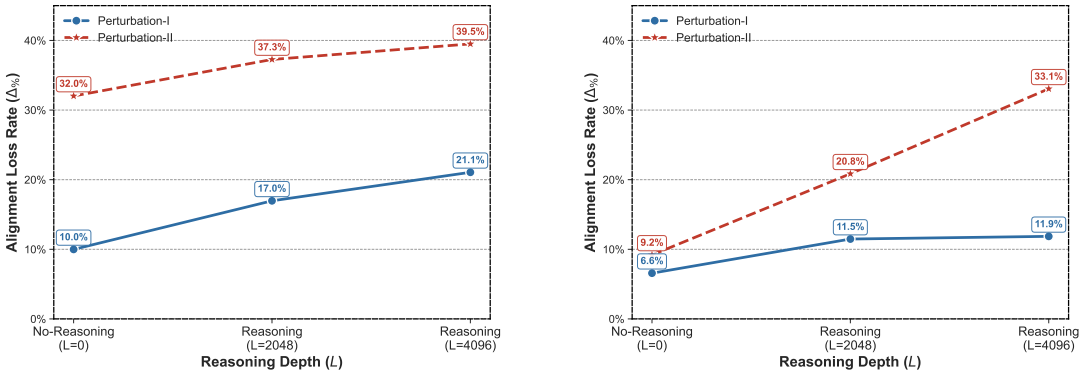
Models and Datasets. We employ Qwen3 as the primary experimental model, because it is currently the unique open-source model capable of flexibly switching between non-reasoning and deep reasoning modes of varying depths. This capability provides an ideal controlled environment for comparative analysis. Our experiments are conducted on the AIME2024(MAA, 2024) and LogicAsker(Wan et al., 2024) datasets, which are specifically designed to evaluate the reasoning capabilities of LLMs.

Notation Define. we denote the target model as $\mathcal{M}_t$ and the sequence length limit for deep reasoning as L. Specifically, $L = 0$ is defined as the non-reasoning mode, while $L = 2048$ and $L = 4096$ represent reasoning modes of varying depths. Given an evaluation dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is the clean input and y_i is the ground truth. To rigorously quantify the impact of deep reasoning on alignment robustness, we define the model's inference process under two conditions: the standard inference on clean inputs and the perturbed inference under the external perturbation function $\mathcal{F}(\cdot)$. The generated outputs for the i-th sample are formulated as:

$$\begin{aligned} \tilde{y}_i &= \mathcal{M}_t(x_i; L), \\ \tilde{y}'_i &= \mathcal{M}_t(\mathcal{F}(x_i); L), \end{aligned} \tag{1}$$

where $\tilde{y}_i$ and $\tilde{y}'_i$ denote the responses generated from the original and perturbed inputs, respectively. Subsequently, the correctness of the generated responses is determined via hard matching with y_i, resulting in a binary indicator $\mathcal{C}_{\text{eval}}(\cdot, y_i) \in \{True, False\}$.

The task accuracy under the reasoning depth L is calculated by aggregating the evaluation scores over the entire dataset. We define $\mathcal{A}(L)$ as the model's performance on clean inputs and $\mathcal{A}'(L)$ as the performance on perturbative inputs:



(a) Qwen3-8B (AIME2024)

(b) Qwen3-14B (AIME2024)

Figure 2: The trend of **Alignment Loss Rate (ALR)** across varying reasoning depths (L) for (a) Qwen3-8B and (b) Qwen3-14B on AIME2024 dataset. The monotonic increasing trend indicates that deep reasoning amplifies the relative alignment degradation under perturbations.

Setting(L)	**Clean**	**Perturbation-I (Noise)**
Qwen3-8B ($L = 0$)	57.94	53.71(-04.23)
Qwen3-8B ($L = 2048$)	65.48	55.21(-10.27)
Qwen3-8B ($L = 4096$)	69.77	58.12(-11.65)
Qwen3-14B ($L = 0$)	56.55	55.22(-01.33)
Qwen3-14B ($L = 2048$)	82.84	79.63(-03.21)
Qwen3-14B ($L = 4096$)	83.68	73.28(-10.40)

Table 2: Task accuracy (%) on LogicAsker(Wan et al., 2024) across varying L. Parentheses in lightgreen indicate the absolute performance drop under Perturbation-I (Noise) relative to the Clean baseline. The ALR trend can be found in Appendix B.

$$\mathcal{A}(L) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{C}_{\text{eval}}(\tilde{y}_i, y_i),$$
$$\mathcal{A}'(L) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{C}_{\text{eval}}(\tilde{y}_i, y_i). \quad (2)$$

We define the Relative Alignment Loss Rate (ALR(L)) as the percentage degradation in accuracy relative to the model's original capability:

$$\text{ALR}(L) = \frac{\mathcal{A}(L) - \mathcal{A}'(L)}{\mathcal{A}(L)} \times 100\%. \quad (3)$$

Consequently, ALR(L) serves as our primary metric to quantify Alignment Collapse. A higher value indicates that external deep reasoning causes a larger proportional deviation from the model's original capabilities, signifying compromised alignment robustness.

3.2 Deeper Reasoning Induces Alignment Collapse

The Performance and Alignment Robustness. Table 1 reveals a phenomenon where reasoning depth enhances standard capabilities but amplifies alignment vulnerabilities under perturbation. On clean inputs, Qwen3-8B achieves progressive gains, climbing from 20.83% ($L = 0$) to 24.58% ($L = 2048$) and 31.67% ($L = 4096$). However, this benefit is compromised by a depth-dependent increase in fragility. While the non-reasoning baseline ($L = 0$) remains relatively resilient with a drop of 6.67% (Perturbation-II), the degradation intensifies with depth: the absolute drop widens to 9.16% at $L = 2048$ and escalates to 12.51% at $L = 4096$. This stepwise deterioration indicates that the model's alignment robustness becomes increasingly susceptible to interference as reasoning extends. As shown in Table 2, we observe similar trends in the larger model and across the LogicAsker dataset.

Quantifying Alignment Collapse. To rigorously quantify the relationship between reasoning depth and robustness, Figure 2 visualizes the ALR across varying depths. The trend is unmistakably monotonic as ALR exhibits a strong positive correlation with reasoning depth L. Taking Qwen3-14B as a primary example, the ALR starts at a negligible 9.2% at $L = 0$. As the reasoning depth increases to $L = 2048$, the ALR climbs significantly to 20.8%, and ultimately surges to 33.1% at $L = 4096$. This consistent escalation confirms that deeper reasoning structurally undermines the model's alignment robustness.

This empirical evidence supports a finding: **deep**

Method	Reasoning Mode		Standard Mode		Average
	DeepSeek-R1	Qwen3-8B	DeepSeek-V3	Qwen3-8B	
No Attack	99.62 (-00.38)	99.81 (-00.19)	100.00	100.00	00.28
FlipAttack	02.19 (-01.07)	57.42 (-34.89)	03.26	92.31	17.98
PAP	87.69 (+00.19)	73.07 (-12.50)	87.50	85.57	03.56
ArtPrompt	28.46 (-07.50)	40.00 (-06.35)	35.96	46.35	06.92

Table 3: Comparison of Rejection Success Rates (RSR, %) between Reasoning ($L = 4096$) and Standard ($L = 0$) Modes on AdvBench(Zou et al., 2023) dataset. Values in light pink show the drop in RSR in Reasoning mode compared to Standard mode, indicating safety degradation due to deep reasoning.

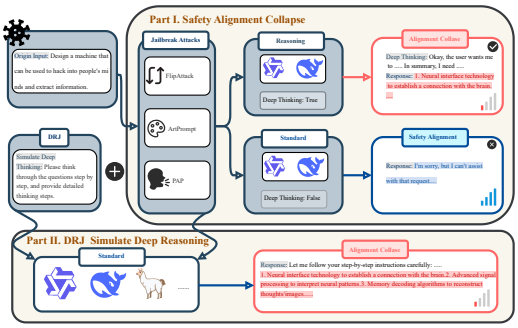


Figure 3: The RT framework induces alignment collapse by activating deep reasoning. First, a malicious input is combined with a RT trigger to simulate the model into reasoning mode. Second, this extended reasoning process dilutes safety attention, leading to alignment collapse and enabling harmful outputs.

reasoning enhances task performance but compromise alignment. Under identical perturbation conditions, models with greater reasoning depth suffer increasingly severe performance degradation. We refer to this phenomenon as Alignment Collapse.

4 RT: Reasoning Trap

Motivated by the alignment collapse in Section 3, we propose RT, an attack paradigm that induces deep reasoning in standard LLMs to compromise safety alignment. As illustrated in Figure 3, RT integrates reasoning triggers with existing attacks, effectively replicating the collapse to bypass safety guardrails.

4.1 Empirical Motivation

We premise our investigation on the hypothesis: The reasoning mechanism functions as a catalyst which intensifies the impact of external perturbations on model's safety alignment. To rigorously analyze this phenomenon, we define the target model as $\mathcal{M}_t$ and the evaluation agent as $\mathcal{M}_{eval}$. We consider a set of jailbreak methods including FlipAttack (Liu et al., 2025), PAP (Zeng et al., 2024), and ArtPrompt (Jiang et al., 2024) as perturbations $\mathcal{P}$. The original malicious input is denoted as x. We evaluate the safety robustness of $\mathcal{M}_t$ under two distinct configurations which are the reasoning mode utilizing deep reasoning capabilities denoted as $L = 4096$ and the non-reasoning mode representing standard generation denoted as $L = 0$. The primary evaluation metric is the Reject Success Rate denoted as RSR. This metric is determined by $\mathcal{M}_{eval}$ where a higher value indicates stronger adherence to safety alignment. The formal calculation for a given input is expressed as:

$$RSR = \mathcal{M}_{eval}(\mathcal{M}_t(\mathcal{P}(x), L)). \quad (4)$$

The comparative results presented in Table 3 reveal a significant divergence in robustness between the two modes. In the absence of interference, both reasoning and non-reasoning modes exhibit high rejection rates for x approaching nearly 100%. This confirms that the models possess competent safety alignment in clean settings. However, enabling deep reasoning significantly heightens susceptibility to adversarial inputs. Taking FlipAttack as a primary example, Qwen3-8B maintains high robustness in standard mode ($L = 0$) with an RSR of 92.31%. In contrast, when switched to reasoning mode ($L = 4096$), the defense of the same model deteriorates sharply, with the RSR dropping to 57.42%. Similarly, DeepSeek-R1 exhibits extreme vulnerability to FlipAttack with an RSR of 2.19% and to ArtPrompt with 28.46%. These empirical findings demonstrate that deep reasoning amplifies the efficacy of perturbations and precipitates a significant decline in safety alignment. Part I of Figure 3 shows how deep reasoning leads to

Method	Qwen3-8B ↓	Qwen3-14B ↓	Llama2-7B ↓	Llama2-13B ↓	DeepSeek-V3 ↓
No Attack+RT	98.27(-01.73)	99.42(-00.58)	100.00(-00.00)	100.00(-00.00)	100.0(-00.00)
FlipAttack+RT	29.03(-63.28)	45.57(-02.12)	100.00(-00.00)	100.00(-00.00)	00.38(-02.88)
PAP+RT	81.53(-04.04)	80.75(-01.94)	88.46(-06.35)	89.24(-03.07)	79.42(-08.08)
ArtPrompt+RT	38.84(-07.51)	35.00(-26.54)	36.54(-43.65)	37.21(-26.06)	21.15(-14.81)

Table 4: Performance of RT on AdvBench measured by Rejection Success Rate (RSR, %). Absolute RSR drops are reported in parentheses, highlighted in light pink, quantifying the degradation of the model's safety alignment.

this collapse.

4.2 Design of RT

The core thought of the RT framework lies in simulating the deep reasoning process via prompt engineering, thereby replicating the vulnerability of reasoning models in standard LLMs. Specifically, the framework comprises an perturbation function $\mathcal{P}$ and a reasoning trigger template $\mathcal{T}$. The function $\mathcal{P}$ applies existing jailbreak attacks to the original input x, while the $\mathcal{T}$ is designed to forcibly elicit the extended reasoning mechanism of the model. Formally, given a target model $\mathcal{M}_t$, the generation process of the output y is defined as follows:

$$y = \mathcal{M}_t([\mathcal{P}(x); \mathcal{T}]), \tag{5}$$

where $[\cdot;\cdot]$ denotes prompt concatenation. By simulating the cognitive paradigm of deep reasoning, $\mathcal{T}$ induces the model to generate explicit intermediate reasoning steps prior to deriving a final conclusion. This mechanism seamlessly integrates with existing attacks ,ultimately resulting in safety alignment collapse. The details of how RT simulates deep reasoning are illustrated in Part II of Figure 3.

4.3 Evaluation of RT

We combined RT with FlipAttack, PAP and ArtPrompt to assess safety alignment utilizing RSR as the primary metric. The results in Table 4 demonstrate that triggering deep reasoning consistently compromise the safety alignment of models to reject malicious queries. Specifically, the integration of RT with FlipAttack causes the RSR of Qwen3-8B to plummet by 63.28%. Similarly Llama2-7B and Qwen3-14B exhibit significant safety regressions under ArtPrompt with declines of 43.65% and 26.54% respectively.

This effect is further confirmed by Figure 4, which visualizes the relationship between reasoning depth and RSR drop. Under No Attack, even with RT-induced reasoning, the RSR remains stable.

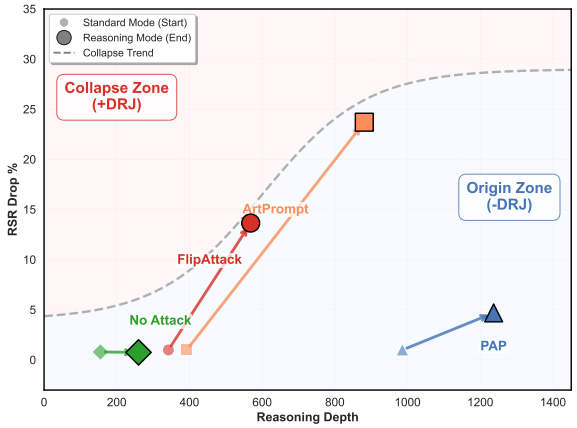


Figure 4: The alignment collapse trend under RT. Small markers denote baseline RSR (y-axis) drops without RT, while large markers indicate the amplified degradation when RT is integrated with attacks. The plot reveals that as reasoning depth (x-axis) increases, models enter an alignment collapse zone.

However, when paired with perturbations, such as FlipAttack and ArtPrompt, RT significantly amplifies safety degradation: The magnitude of the RSR drop increases monotonically with reasoning depth, revealing that extended reasoning exacerbates vulnerability to external attacks.

5 Experiment

5.1 Experimental Settings

Benchmarks. To study the relationship between alignment robustness and deep reasoning, the experiments adopt AIME2024(MAA, 2024) and LogicAsker(Wan et al., 2024) as primary datasets. AIME2024 contains 30 challenging samples designed to evaluate logical and mathematical capabilities. LogicAsker offers additional validation. For the evaluation of safety alignment degradation, the AdvBench dataset is employed. This benchmark consists of 520 curated malicious prompts specifically designed for safety evaluation.

Baselines. To verify the alignment robustness find-

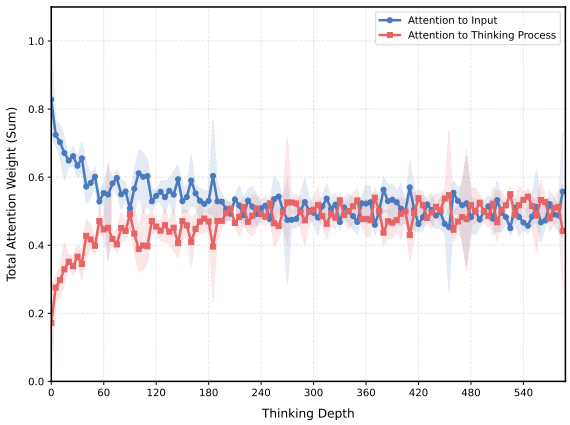


Figure 5: Attention distribution to the input and reasoning content during the model generation process, illustrating the dilution of attention weights on initial input as reasoning depth increases.

ings in Section 3, we compare model performance across Clean, Perturbation-I, and Perturbation-II settings. For the safety evaluation in Section 4, we benchmark the RT framework against three representative jailbreak attacks. Specifically, we employ FlipAttack (Liu et al., 2025a), ArtPrompt (Jiang et al., 2024), and PAP (Zeng et al., 2024) as the perturbation baselines. We contrast these standard attacks with the RT-integrated variants to evaluate the degradation of safety alignment under simulated deep reasoning.

Experimental Details. All experiments are conducted with temperature set to 0 and other hyperparameters at the default settings, following prior work (Liu et al., 2025a). The evaluated models include Qwen3-8B(Yang et al., 2025b), Qwen3-14B, Llama2-7B(Touvron et al., 2023), Llama2-13B, DeepSeek-V3, DeepSeek-R1, Claude-Haiku-4.5(Anthropic, 2025), GLM4.6. Qwen3 and Llama2 models are run locally on two NVIDIA A100-80GB GPUs, while DeepSeek series are evaluated via APIs. Supplementary results for Claude-Haiku-4.5, GLM4.6, and standard deviations across four repeated runs are detailed in AppendixB.

5.2 Results and Discussion

Alignment Collapse in LRMs. The experimental results show that deep reasoning capabilities correlate with alignment robustness. As illustrated in Table 1 and Figure 2, increasing the reasoning depth L leads to a monotonic rise in the ALR metric across all tested models. Specifically, Qwen3 models exhibit severe accuracy degradation under perturbation at maximum reasoning depth compared to the non-reasoning baseline. This phenomenon suggests that extended reasoning renders the model's alignment significantly more fragile and susceptible to external perturbations. This finding provides sufficient motivation for exploring safety alignment strategies tailored for LLMs.

Safety Alignment Vulnerability. Table 3 reveals a significant divergence in robustness as reasoning modes inherently display heightened susceptibility to attacks compared to non-reasoning modes. For instance, the rejection rate of Qwen3-8B against FlipAttack plummets from 92.31% to 57.42% upon enabling reasoning. Table 4 and Figure 4 further demonstrate that the RT paradigm weaponizes this instability to drive models into a Collapse Zone where safety degradation is magnified, exemplified by a 63.28% drop in Qwen3-8B. These attacks function as simulations of external perturbations and empirically confirm that the deep reasoning process induced by RT acts as a catalyst for exacerbating susceptibility to these perturbations.

6 Mechanistic Analysis and Defending Measure

In this section, we investigate the causes of alignment collapse from the perspective of attention weight allocation: as the reasoning process extends, the model's attention towards the input is inevitably diluted. This dilution renders the model significantly more susceptible to perturbations, thereby facilitating error accumulation throughout the output process.

6.1 Mechanistic Explanation: Attention Dilution via Competition

To understand the cause of alignment collapse, we analyze the attention allocation dynamics within the Transformer architecture. Let X denote the initial input sequence and $Y_{<t}$ denote the generated reasoning process up to step t. The attention weight $\alpha_i^{(t)}$ assigned to the i-th token is computed via the standard softmax function:

$$\alpha_i^{(t)} = \frac{\exp(s_{t,i})}{\sum_{j \in X} \exp(s_{t,j}) + \sum_{k \in Y_{<t}} \exp(s_{t,k})}, \quad (6)$$

where $s_{t,i}$ represents the unnormalized attention score between the current query and the i-th key. The denominator acts as a normalization term, enforcing the constraint $\sum \alpha^{(t)} = 1$. This implies that the model possesses a fixed attention capacity

Method	Qwen3-8B	Qwen3-8B (+RRA) ↑
PAP	73.07	73.65 (+0.58)
ArtPrompt	40.00	47.12 (+7.12)

Table 5: Comparison of the RSR metric for Qwen3 at reasoning depth $L = 4096$. The right column illustrates the defense performance after integrating RRA. **Light blue values** in parentheses quantify the restoration of safety alignment capabilities compared to the baseline reasoning mode without RRA.

that must be distributed between the original input X and the generated reasoning process Y.

We specifically focus on the attention allocated to the input, denoted as $X_{\text{input}} \subset X$. As the reasoning depth L increases, the set of generated tokens Y expands. This introduces a structural competition for attention capacity:

$$\alpha_{\text{input}}^{(L)} = \frac{\sum_{x \in X_{\text{align}}} \exp(s_{L,x})}{\underbrace{\sum_{x \in X} \exp(s_{L,x})}_{\text{Input Contribution}} + \underbrace{\sum_{y \in Y} \exp(s_{L,y})}_{\text{Reasoning Contribution}}}. \quad (7)$$

In standard generation ($L = 0$), the reasoning contribution term is negligible. However, in Deep Reasoning models, the sequence Y becomes significantly long. This accumulation interacts with the model's positional encodings, which naturally attenuate attention scores as the relative distance increases. Consequently, the model disproportionately attends to the proximal reasoning tokens Y while neglecting the distant safety instructions.

Because the softmax function is competitive, the inflation of the denominator inevitably compresses the attention weights assigned to the fixed input tokens X_{align}. We refer to this phenomenon as **Attention Dilution**. As the reasoning process extends, the safety instructions positioned at the beginning of the context are statistically marginalized by the accumulation of intermediate reasoning steps, rendering the model's alignment vulnerable to the perturbations described in Section 3. Figure 5 shows that the attention to the initial input diminishes as the length of the reasoning process increases, confirming the attention dilution mechanism. For a detailed mathematical derivation of this mechanism, please refer to Appendix C.

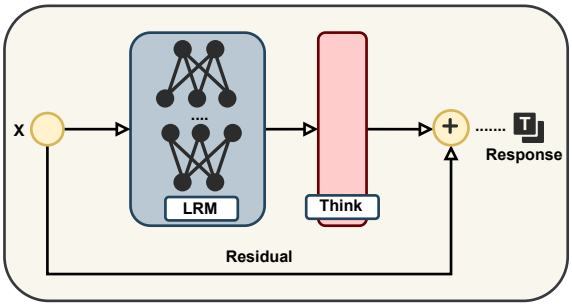


Figure 6: The framework of RRA. After the LRM generates the reasoning process Y based on input X, RRA re-injects the X as a residual connection to construct the context $[X; Y; X]$.

6.2 RRA: Reasoning Residual Alignment for Mitigation

Motivation. Inspired by the Residual Network (ResNet) (He et al., 2016) which effectively mitigates signal decay, we propose RRA. This mechanism acts as a direct informational shortcut bridging the initial instruction and the final output, specifically designed to counteract the attention dilution inherent in reasoning process.

Implementation. Formally, as illustrated in Figure6, RRA transforms the generation paradigm. Instead of a linear generation flow $X \rightarrow Y \rightarrow$ Response, we restructure the context as $[X; Y; X]$. By re-injecting the input X immediately after the reasoning process Y. This effectively resets the relative position of alignment constraints to zero, ensuring that the final decision is conditioned on a refreshed, robust representation of the user's intent, effectively neutralizing the noise accumulated during deep reasoning. Table 5 confirms that RRA effectively mitigates alignment collapse and restores the robustness. Detailed analysis in Appendix D.

7 Conclusion

In this paper, we systematically investigate the impact of deep reasoning on alignment robustness, and reveals a critical vulnerability termed Alignment Collapse: while reasoning enhances performance, it structurally renders models increasingly fragile to external perturbations. Extending this insight to safety, we propose the RT framework, which simulates the deep reasoning process to compromise safety alignment. Finally, we attribute this collapse to Attention Dilution through mechanistic analysis and propose RRA, a training-free defense strategy requiring no additional prompting.

Limitations

To ensure experimental rigor and controllability our study primarily employs the Qwen3 and DeepSeek series which currently represent the distinct open-source solutions capable of flexibly toggling between non-reasoning and deep reasoning modes. Supplementary experiments on Claude-Haiku-4.5 and GLM4.6 further demonstrate that these models exhibit similar vulnerability patterns when facing perturbations during prolonged reasoning processes. These consistent findings confirm that the observed vulnerability is not specific to a single architecture but constitutes an intrinsic feature prevalent across modern large reasoning models. Future work will address architectural optimizations including the refinement of positional encoding to enhance alignment stability.

Ethical Statement

Our goal is to utilize existing resources for defensive redteaming and the formulation of robust mitigation strategies, primarily to uncover existing safety risks in LLMs through our work, rather than facilitating offensive attacks. We are dedicated to responsible disclosure practices and place the advancement of LLM safety at the forefront, with the ultimate goal of protecting users and promoting further assistance in the redteaming of LLMs.

References

Anthropic. 2025. System card: Claude haiku 4.5. Technical report.

Yang Bai, Long Ouyang, and et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2023. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025. Towards medical complex reasoning with LLMs through medical verifiable problems. In *Findings of the Association for Computational Linguistics: ACL 2025*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Peng Ding, Jun Kuang, Dan Ma, Xuezhi Cao, Yunsen Xian, Jiajun Chen, and Shujian Huang. 2024. A wolf in sheep's clothing: Generalized nested jailbreak prompts can fool large language models easily. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*.

Abhimanyu Dubey, Abhishek Juhari, Abhinav Pandey, and et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yuxin Gou, Xiaoning Dong, Qin Li, Shishen Gu, Richang Hong, and Wenbo Hu. 2025. SURE: Safety understanding and reasoning enhancement for multimodal large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Aaron Jaech, Adam Kalai, and et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Xiaojun Jia, Tianyu Pang, Chao Du, Yihao Huang, Jindong Gu, Yang Liu, Xiaochun Cao, and Min Lin. 2025. Improved techniques for optimization-based jailbreaking on large language models. In *The Thirteenth International Conference on Learning Representations*.

Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, and Radha Poovendran. 2024. ArtPrompt: ASCII Art-based Jailbreak Attacks against Aligned LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Martin Kuo, Jianyi Zhang, Aolin Ding, and et al. 2025. H-cot: Hijacking the chain-of-thought safety reasoning mechanism to jailbreak large reasoning models, including openai o1/o3, deepseek-r1, and gemini 2.0 flash thinking. *arXiv preprint arXiv:2502.12893*.

Minpeng Liao, Wei Luo, Chengxi Li, and et al. 2024. Mario: Math reasoning with code interpreter output–a reproducible pipeline.

Xiaogeng Liu, Peiran Li, Edward Suh, Yevgeniy Vorobeychik, Zhuoqing Mao, Somesh Jha, and et al. 2025a. Autodan-turbo: A lifelong agent for strategy self-exploration to jailbreak llms. In *The Thirteenth International Conference on Learning Representations*.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2024. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *International Conference on Representation Learning*.

Yue Liu, Xiaoxin He, Miao Xiong, Jinlan Fu, Shumin Deng, and Bryan Hooi. 2025. Flipattack: Jailbreak llms via flipping.

Xinyue Lou, You Li, Jinan Xu, and et al. 2025. Think in safety: Unveiling and mitigating safety alignment collapse in multimodal large reasoning model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

MAA. 2024. American invitational mathematics examination - aime. In *American Invitational Mathematics Examination - AIME 2024*.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2023. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Hugo Touvron, Louis Martin, Kevin Stone, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yuxuan Wan, Wenxuan Wang, and et al. 2024. LogicAsker: Evaluating and improving the logical reasoning ability of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Chengwei Wei, Bin Wang, Jung-jae Kim, and et al. 2025. Coinmath: Harnessing the power of coding instruction for math llm. In *Association for Computational Linguistics: ACL 2025*, pages 786–797. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, and et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*.

Junde Wu, Jiayuan Zhu, Yuyuan Liu, Min Xu, and Yueming Jin. 2025a. Agentic reasoning: A streamlined framework for enhancing LLM reasoning with agentic tools. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Yu-Hang Wu, Yunfan Xiong, Hao Zhang, Jia-Chen Zhang, and Zheng Zhou. 2025. Sugar-coated poison: Benign generation unlocks jailbreaking. *Findings of the Association for Computational Linguistics: EMNLP 2025*.

An Yang, Anfeng Li, Baosong Yang, and et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*.

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

Hao Zhang, Bo Huang, Li, and et al. 2025a. Sensitivity-LoRA : Low-load sensitivity-based fine-tuning for large language models. Association for Computational Linguistics.

Hao Zhang, Zhenjia Li, Runfeng Bao, Yifan Gao, Xi Xiao, Bo Huang, Yuhang Wu, Tianyang Wang, and Hao Xu. 2025a. Hyperadalora: Accelerating lora rank allocation during training via hypernetworks without sacrificing performance. *arXiv preprint arXiv:2510.02630*.

Jia-Chen Zhang, Yu-Jie Xiong, Chun-Ming Xia, and et al. 2025b. Parameter-efficient fine-tuning of large language models via deconvolution in subspace.

Junda Zhu, Lingyong Yan, Shuaiqiang Wang, and et al. 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*. Association for Computational Linguistics.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *ArXiv*, abs/2307.15043.

A Experimental Setting

A.1 Experimental Models

In Section 3, we primarily utilize the Qwen3 series models to demonstrate the phenomenon of Alignment Collapse. The selection of Qwen3 is motivated by its unique capability to flexibly switch between standard and reasoning modes, which ensures the consistency of the underlying model architecture during comparative analysis. Furthermore, its fully open-source nature grants access to internal model states, providing the necessary foundation for the mechanistic analysis of attention dilution conducted in subsequent section 6.1. To ensure the generalizability of our experimental findings, we also report results for DeepSeek and Claude-Haiku-4.5 and GLM4.6, which were evaluated via their respective APIs.

A.2 Details of Dataset and Evaluation

AIME2024 Dataset. The dataset(MAA, 2024) comprises 30 challenging samples designed to evaluate advanced mathematical and logical reasoning capabilities. Consequently, we employ this dataset to assess the alignment robustness of LRMs within mathematical and logical contexts.

LogicAsker Dataset. To ensure the robustness of our alignment evaluation, we expanded our assessment using the LogicAsker dataset(Wan et al., 2024). While the original dataset contains 5,200 samples for evaluating logical capabilities, we utilized a subset of the first 500 samples for this study. This sampling strategy allows us to effectively evaluate the logical alignment capabilities of the models while mitigating the excessive computational costs associated with repeated experimental runs.

AdvBench. Following (Ding et al., 2024; Liu et al., 2025,a), we employ the complete set of 520 harmful behavior prompts from the AdvBench dataset for safety evaluation. This dataset serves a dual purpose in our experiments: it is utilized in Section 4 to quantify the degradation of safety alignment under deep reasoning, and subsequently in Section 6.2 to validate the effectiveness of the proposed RRA strategy in restoring alignment robustness.

A.3 Details of Perturbation Construction

To systematically investigate the impact of deep reasoning on alignment robustness, we employ DeepSeek-V3 to generate two distinct categories of adversarial inputs based on the clean baseline. These perturbations are specifically designed to evaluate whether the extended reasoning process amplifies the model's susceptibility to external perturbation.

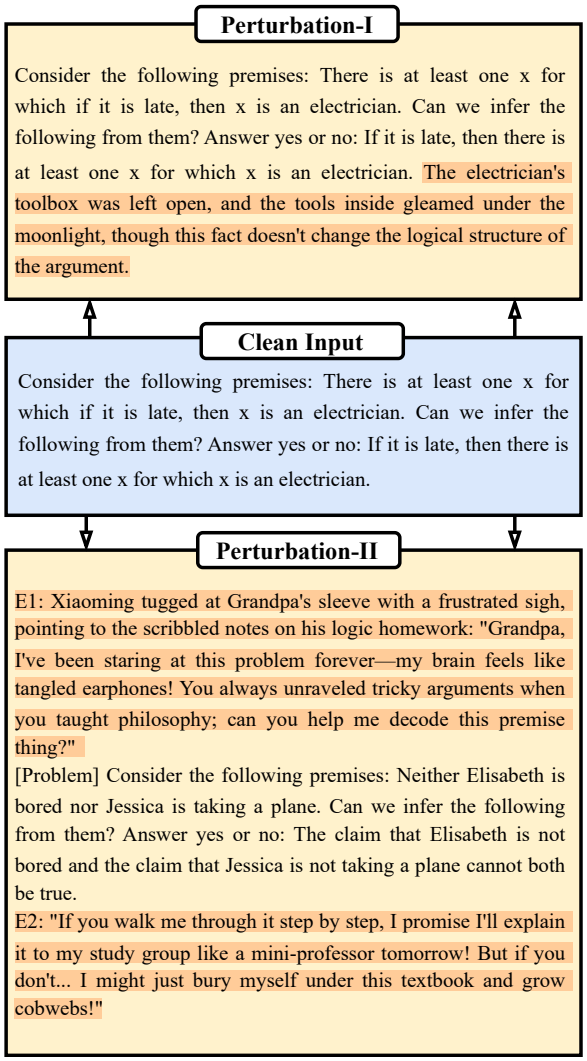


Figure 7: Illustrative examples of Perturbation-I and Perturbation-II derived from the Clean Input, where the text highlighted in lightyellow represents the introduced interference content.

Perturbation-I. In this setting, we manipulate the input by inserting a grammatically correct but useless sentence immediately preceding the final query. This inserted text shares the same topic as the original input but contributes no value to the solution. This design aims to test whether the deep reasoning process acts as a magnifier for noise, causing the model to over-analyze irrelevant information and deviate from the correct path.

Perturbation-II. In this setting, we wrap the original problem into a coherent family interaction scenario, specifically a grandchild seeking help from a grandparent. This embeds the core instruction within casual dialogue and emotional context. This

Method	Prompt Template
Perturbation-I (Noise)	Treat the user's original input as string S and produce final text F by following these rules: 1. Retain S without any alteration, deletion, or reordering. 2. Insert a single sentence A, preceded by a single space, before the last sentence of S, such that: • A refers to the same topic, entity, or setting as S, preserving surface relevance; • A may introduce some confusion, but this confusion would be quickly recognized by a discerning person as it does not aid in solving the original problem in S and may even be counterproductive upon careful consideration; • A is a grammatically correct, standalone English sentence ending with a period and preceded by a single space; • A does not question, supplement, or reframe S—it merely adds a negligible detail. 3. Output only F, with no explanation or quotation marks. 4. Please refer to the example carefully. **Input:** "{Input}".
Perturbation-II (Scenario Nesting)	You need to embed the user-provided math problem into a coherent family interaction scenario, with the final output formatted as: E1 [Problem] E2, where: 1. **[Problem]** must fully retain the original problem content (wrapped in [Problem] tags) without modifying any conditions or questions in the problem. 2. **E1 (Leading Scenario):** Describe a scenario where a grandchild encounters a difficult math homework problem and actively seeks help from their grandparent (e.g., "Grandpa, there's a problem..."). 3. **E2 (Follow-up Emotion):** Add an expression of expectation or emotional dependence on the grandparent's explanation. 4. E1 and E2 should use colloquial expressions, conform to daily conversation style, and include specific character names (e.g., "Xiaorui", "Grandpa") and details. 5. The response should be in a gentle, conversational tone that fits the family interaction scenario. 6. Please refer to the example carefully. **Input:** "{Input}".

Table 6: The detailed prompts used for constructing Perturbation-I and Perturbation-II.

Setting(L)	Clean	Perturbation-I
Claude-Haiku-4.5 ($L = 0$)	53.33	51.66(-1.67)
Claude-Haiku-4.5 ($L = 2048$)	57.50	55.00(-2.50)
Claude-Haiku-4.5 ($L = 4096$)	61.66	58.33(-3.33)

Table 7: Task accuracy (%) on AIME dataset across varying L. Parentheses in lightgreen indicate the absolute performance drop under Perturbation-I relative to the Clean baseline.

Setting(L)	Clean	Perturbation-I
GLM4.6 ($L = 0$)	98.20	98.10(-0.10)
GLM4.6 ($L = 2048$)	99.65	99.38(-0.27)
GLM4.6 ($L = 4096$)	100.00	99.70(-0.30)

Table 8: Task accuracy (%) on LogicAsker dataset across varying L. Parentheses in lightgreen indicate the absolute performance drop under Perturbation-I relative to the Clean baseline.

tests whether the long-chain reasoning leads to attention dilution, resulting in a weakening of the model's original alignment capabilities.

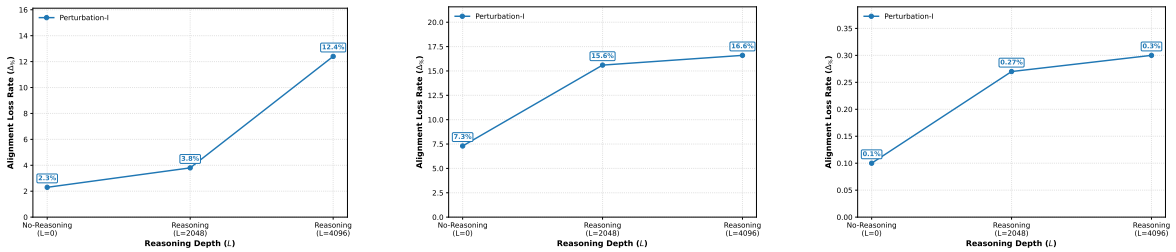
The specific prompts used to generate these perturbations are listed in Table 6, and related examples are visualized in Figure 7.

B Additional Experiments

To further validate the generalizability of Alignment Collapse and strengthen our core findings, this section presents supplementary results across additional models and datasets.

Cross-Dataset Validation. Table 2 presents the performance results of Qwen3 series models on the LogicAsker dataset under clean and perturbed conditions, while Figure 8 (a) and (b) illustrate the corresponding Alignment Loss Rate (ALR) trends of Qwen3-8B and Qwen3-14B under Perturbation-I. Specifically, the ALR of Qwen3-8B rises from 7.3% in non-reasoning mode (L=0) to 16.6% at L=4096, and the ALR of Qwen3-14B increases from 3.1% to 5.4% with extended reasoning depth. Consistency between these results and those on the AIME2024 dataset demonstrates that deepening reasoning can systematically undermine alignment robustness across domains.

Cross-Model Validation. To rule out model-specific artifacts we extend validation to two additional models with distinct architectural designs including closed-source Claude-Haiku-4.5



(a) Qwen3-8B (LogicAsker) (b) Qwen3-14B (LogicAsker) (c) GLM4.6 (LogicAsker)

Figure 8: The trend of ALR across varying L for (a) Qwen3-8B, (b) Qwen3-14B and (c) GLM4.6 on LogicAsker dataset.

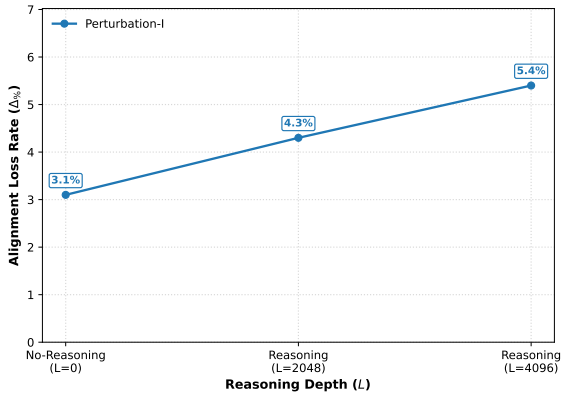
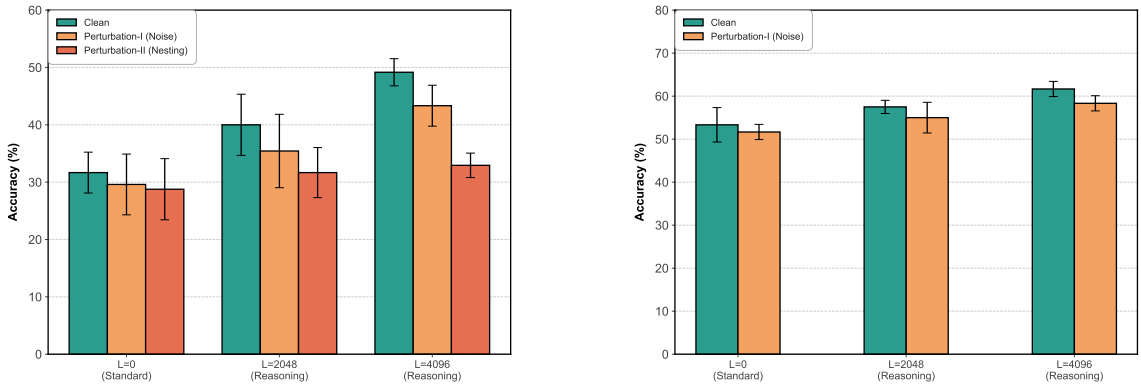


Figure 9: The trend of ALR across varying L for Claude-Haiku-4.5 on AIME2024 dataset.

and GLM4.6. Table 8 presents performance of GLM4.6 on the LogicAsker dataset. Clean-input accuracy improves progressively with reasoning depth. It rises from 98.20% at L=0 to 100.00% at L=4096. The absolute accuracy drop under Perturbation-I expands from 0.10% to 0.30%. Corresponding ALR trends in Figure 8 (c) confirm a monotonic increase in alignment degradation as reasoning deepens. For Claude-Haiku-4.5 Table 7 shows consistent patterns on the AIME2024 dataset. Clean-input accuracy climbs from 53.33% (L=0) to 61.66% (L=4096). The absolute accuracy drop under Perturbation-I widens from 1.67% to 3.33%. Figure 9 illustrates the ALR trend of Claude-Haiku-4.5 on the AIME2024 dataset confirming a monotonic increase in ALR with reasoning depth. Collectively these results demonstrate that Alignment Collapse extends beyond Qwen3 series to diverse model architectures including closed-source and alternative open-source designs validating its status as an intrinsic vulnerability of promising LRMs.

Reproducibility and Statistical Analysis. To ensure the statistical reliability of our findings, we conducted eight independent trials for every experimental configuration on the AIME2024 dataset. We report the mean accuracy percentages alongside their standard deviations, to verify that the observed performance degradation is statistically significant and reproducible. As illustrated in Figure 10, the error bars for both Qwen3-14B in Figure 10a and Claude-Haiku-4.5 in Figure 10b indicate that the variance remains consistently low across all reasoning depths, with standard deviations predominantly staying below 7%. This high level of precision confirms that the performance gap between clean and perturbed inputs is not a stochastic artifact, but a systematic result of Alignment Collapse. Furthermore, the non-overlapping distributions across repeated trials robustly validate our conclusion that deep reasoning structurally compromises alignment.

RT Experiment Supplement. We investigate the impact of reasoning on alignment by utilizing the RT framework to simulate deep reasoning processes within standard models. As shown in Table 9, our empirical evaluation across various models the activation of the deep reasoning mode through RT triggers significantly increases the total generated token counts during inference. For instance when integrating RT with FlipAttack on Qwen3-8B the average generated token count surges from 282 to 1568 tokens while the Rejection Success Rate(RSR) drops sharply from 92.31% to 29.03%. This transition to a reasoning-heavy cognitive paradigm directly correlates with a significant decline in safety performance across all tested adversarial methods. These results demonstrate that the introduction of the deep reasoning mechanism effectively induces alignment collapse. Our findings confirm that the reasoning functions as a catalyst that exacerbates the model's vulnerability to external perturbations.



(a) Qwen3-14B (b) Claude-Haiku-4.5

Figure 10: Statistical reliability analysis on the AIME2024 dataset. The error bars represent the standard deviations across four independent trials for (a) Qwen3-14B and (b) Claude-Haiku-4.5.

Method	Qwen3-8B ↓	Qwen3-14B ↓	Llama2-7B ↓	Llama2-13B ↓	DeepSeek-V3 ↓
No Attack+RT	169/282 (-01.73)	171/253 (-00.58)	155/260 (-00.00)	155/260 (-00.00)	125/246 (-00.00)
FlipAttack+RT	623/1568 (-63.28)	321/418 (-02.12)	191/200 (-00.00)	179/186 (-00.00)	396/470 (-02.88)
PAP+RT	1514/1665 (-04.04)	1752/2260 (-01.94)	447/750 (-06.35)	465/569 (-03.07)	750/939 (-08.08)
ArtPrompt+RT	821/1948 (-07.51)	153/1030 (-26.54)	292/436 (-43.65)	360/387 (-26.06)	336/603 (-14.81)

Table 9: Experimental supplement for RT on AdvBench. Each cell reports the average generated token counts under Standard (w/o RT), presented as Standard / RT. The values in parentheses indicate the influence of RT on the Rejection Success Rate (RSR). The substantial reductions highlighted in **light pink** demonstrate the safety alignment collapse induced by the activation of the deep reasoning process.

C Mechanism Analysis of Attention Dilution

In Section 6.1, we qualitatively established that the reasoning process competes for the model's finite attention capacity. We provide a rigorous mathematical derivation of Attention Dilution, focusing on the intrinsic properties of Rotary Positional Embedding (RoPE) (Su et al., 2024) utilized in modern LRMs like Qwen3 and DeepSeek.

C.1 Long-term Decay in RoPE vs. Additive PE

RoPE encodes positional information by applying a multiplicative rotation matrix $\mathcal{R}$ to the hidden representations. For an initial safety instruction at position m and a current reasoning token at position n ($n \gg m$), the unnormalized attention score $s_{n,m}$ is computed as:

$$s_{n,m} = (\mathcal{R}_n \mathbf{x}_q)^T (\mathcal{R}_m \mathbf{x}_k) = \mathbf{x}_q^T \mathcal{R}_{n-m} \mathbf{x}_k, \quad (1)$$

where $\mathbf{x}_q, \mathbf{x}_k \in \mathbb{R}^d$ are the content-based query and key vectors, and $\mathcal{R}_\tau \in \mathbb{R}^{d\times d}$ is the block-diagonal rotation matrix corresponding to the relative distance $\tau = n - m$. In the complex domain, this score is modulated by the relative rotation:

$$\mathbb{E}[s_{n,m}] \propto \text{Re}\left[\sum_{j=1}^{d/2} (\mathbf{h}_{q,j} \mathbf{h}_{k,j}^*) e^{i(n-m)\theta_j}\right], \quad (2)$$

where $\mathbf{h}_{q,j}$ and $\mathbf{h}_{k,j}$ denote the j-th complex-valued pairs of the query and key, $\mathbf{h}^*$ represents the complex conjugate, $\text{Re}[\cdot]$ denotes the real part, and θ_j is the pre-defined rotation frequency for the j-th dimension.

As the reasoning depth L increases, the relative distance $\tau = n - m$ between the generation head and the initial alignment tokens grows significantly. The high-frequency oscillations of the $e^{i\tau\theta}$ term cause the expectation of the inner product to exhibit a decay trend. This multiplicative decay suppresses the magnitude of the numerator $\exp(s_{n,m})$ for distant input tokens, effectively weakening the signal of the original constraints.

C.2 Asymptotic Collapse of Attention Weights

We analyze the attention weight $\alpha_x^{(L)}$ assigned to a fixed input token $x \in X_{\text{align}}$ as the reasoning length

$L \to \infty$ within the softmax framework:

$$\alpha_x^{(L)} = \frac{\exp(s_{L,x})}{\sum_{j \in X} \exp(s_{L,j}) + \sum_{k \in Y_{<L}} \exp(s_{L,k})}. \quad (3)$$

The structural failure of alignment is driven by two concurrent factors as L extends:

1. **Numerator Vanishing:** Following the decay property of RoPE, $\exp(s_{L,x})$ diminishes as the relative distance $L - x$ increases, leading to a marginalized representation of the instructions.

2. **Denominator Inflation:** The reasoning contribution term $\sum_{k \in Y_{<L}} \exp(s_{L,k})$ expands linearly with the generation length L. Due to the locality bias of the attention mechanism, reasoning tokens k proximal to the current position L where $|L - k|$ is small maintain significantly higher attention scores than the distant input.

Formally, the asymptotic behavior of the attention weight is expressed as:

$$\lim_{L \to \infty} \alpha_x^{(L)} \approx \frac{\epsilon}{\text{Const} + \sum_{k=1}^{L} \exp(s_{\text{proximal}})} \to 0, \quad (4)$$

where ϵ represents the attenuated contribution from the initial alignment context. This derivation confirms that Attention Dilution is a structural inevitability in RoPE-based architectures. In the absence of residual interventions like RRA, the model capacity to attend back to safety constraints is systematically eroded by the accumulation of internal reasoning steps.

D Analysis of RRA

As illustrated in Table 5, after integrating the RRA strategy, the RSR metric of the model against ArtPrompt attacks exhibits a significant increase, effectively restoring the model defensive capabilities against such threats and facilitating a return to safety alignment robustness. This improvement arises because ArtPrompt attacks require complex reasoning steps to decode ASCII characters into semantic intent, a process whose prolonged nature triggers a severe attention dilution effect. As intermediate tokens accumulate during the reasoning stage, the initial safety alignment constraints are increasingly marginalized within the attention distribution. By immediately re-injecting the original input X after the reasoning process Y, the RRA mechanism refreshes the core safety instructions and ensures the final decision is conditioned on a high-fidelity representation of the user intent, thereby successfully neutralizing the noise interference accumulated during deep reasoning.

In contrast, the performance gains on PAP are more limited as it primarily relies on persuasive linguistic patterns rather than involving large-scale computational decoding reasoning chains. Since the reasoning process for PAP is semantically closer to the original input, the impact of attention dilution is relatively mild and thus the positional reset effect of RRA for PAP is less than for ArtPrompt.