# Do Thinking Tokens Help with Safety?

**Anonymous Authors**[1]

## Abstract

Today's reasoning models use thinking tokens to attain stronger performance on benchmarks than their instruction-tuned counterparts. It is also generally believed that this more "deliberative" mode should improve alignment and safety, by providing the model a safe space to consider whether its planned answer to the request violates its safety principles. We present evidence that this intuition is not always correct. Across frontier open-weight reasoning models including GPT-OSS, Phi, OLMo, and Qwen, we find that the model's decision is already strongly encoded at the beginning of thinking, with a probe on the first token's hidden representation predicting refusal/compliance with $\geq 0.85$ AUROC and $\sim 90\%$ balanced accuracy. We also find little evidence of genuine safety deliberation in thinking models, as additional thinking after the first 20% of the trace rarely moves the final decision. While sentence-level inspection of thinking traces show signs of oscillation between refusal- and compliance-leaning rationales, we find that such oscillations exert limited to no influence on the final response in $\geq 85\%$ of thinking traces. Furthermore, existing inference-time and training-based safety interventions largely shift models toward refusal-leaning reasoning, substantially reducing helpfulness on benign prompts. Together, our results suggest that safety behavior in current reasoning models is far less deliberative than assumed, highlighting the need for training methods that use thinking traces effectively for safety-critical decisions.

## 1. Introduction

Recent large reasoning models (LRMs) are trained to utilize inference-time compute for improving general capabilities by generating long reasoning traces before producing a final response, also commonly referred to as "thinking." This use of inference-time compute has led to substantial gains on many verifiable tasks (Wei et al., 2022; Kojima et al., 2022; Zhou et al., 2023; Shao et al., 2024; DeepSeek-AI, 2025), with evidence that reasoning can also be adapted to domains without directly verifiable rewards (Tang et al., 2025; Yu et al., 2025; Ren et al., 2025; Huan et al., 2025).

An important line of research asks whether this reasoning capability can improve safety in the same way that it improves performance (Wang et al., 2025a; Zhou et al., 2025a). Safety is commonly evaluated with two metrics that are often in tension: *Attack Success Rate* (ASR), the rate at which a model complies with harmful prompts, and *Over-Refusal Rate* (ORR), the rate at which a model incorrectly refuses benign prompts. A safe and helpful model should simultaneously attain low ASR and ORR (Kim et al., 2025). A natural hypothesis is that thinking should improve this trade-off, as models may have additional opportunity to recognize harmful requests while preserving helpfulness on benign inputs by deliberating before answering. This intuition is appealing and also underlies several prominent rule- and specification-based safety defenses (Bai et al., 2022; Guan et al., 2024).

Yet, there are strong reasons to question this hypothesis. Recent work (Qi et al., 2025; Liu et al., 2026; Zhao et al., 2025c) shows that safety alignment in instruction-tuned models is often shallow, with behavior largely determined by a small number of initial tokens and thus easily circumvented through simple early token manipulations. If this phenomenon persists in reasoning models, then additional inference-time compute may not enable meaningful deliberation, but instead simply serve to rationalize an already-formed decision. Thus, we ask:

> *Does thinking truly improve safety decisions in reasoning models, and does longer thinking lead to better ASR–ORR tradeoffs?*

As we detail later, our findings confirm this concern: safety decisions in reasoning models remain shallow, and longer thinking does not meaningfully improve safety behavior. But one might hope that defenses tailored to reasoning models address this failure by making the trace more safety-aware, either through inference-time interventions or train-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

ing. This raises a second question:

> *Do existing inference- and training-based defenses meaningfully improve the ASR–ORR trade-off in reasoning models?*

**Our Contributions.** In this work, we provide a detailed study showing that safety behavior in current LRMs may be far less deliberative than the reasoning paradigm suggests. First, we present a surprising finding that models already appear to decide whether to refuse or comply even before any thinking is generated (§2.2). Across four frontier open-weight models, we find that a linear probe on the first token's representation in the thinking trace separates final refusals from compliances with AUROC 0.85–0.95. This probe also matches the model's final response with up to $90\%$ balanced accuracy, indicating that the eventual refusal/compliance outcome is already strongly reflected in the early representation. Notably, the identity of the generated token itself is far less predictive, highlighting that this signal is encoded in the hidden representation rather than the surface token.

Second, we investigate the role of thinking in the safety decision process and find that its impact is limited. Namely, more thinking does not meaningfully translate to improved safety decisions (§2.3), and while thinking traces often exhibit intermediate sentence-level patterns that resemble safety deliberation, these trace segments turn out to have minimal influence on the final response (§2.4).

Third, we study how existing defenses change the reasoning behavior of LRMs by evaluating several inference-time (Jeung et al., 2025; Phan et al., 2025; Kim et al., 2026) and training-based (Wang et al., 2025b; In et al., 2025; Zhou et al., 2025b; Lee et al., 2026; Zhang et al., 2025b; Wei et al., 2026) methods. We find that inference-time methods either have limited impact or reduce ASR at the cost of higher ORR (§3.1). In contrast, training-based methods induce larger behavioral shifts, but do so by pushing models toward generating more refusal-leaning traces on both harmful and benign prompts, again trading lower ASR for substantially higher ORR (§3.2). These findings highlight the need for safety objectives that make thinking meaningfully inform refusal/compliance decisions.

## 2. Safety Behavior of Reasoning Models

Here, we investigate when the refusal/compliance decision emerges during a model's generation and how thinking contributes to the final response. We first show in §2.2 that the representation of the first token in the thinking trace strongly encodes a refusal/compliance signal that corresponds closely to the model's final safety decision. This motivates the question of whether the subsequent thinking trace can alter this early signal. In §2.3, we answer this question in the negative, finding that longer thinking does not lead to more meaningful safety deliberation. In §2.4, we further find that individual trace segments may appear deliberative but have minimal influence on the model's final response.

| | Dataset | # Examples |
|---|---|---|
| **ASR** | WildJailbreak (Jiang et al., 2024) | 2,000 |
| | FORTRESS (Knight et al., 2025) | 500 |
| **ORR** | OR-Bench-Hard (Cui et al., 2025) | 1,319 |
| | FalseReject (Zhang et al., 2025c) | 1,187 |
| | CoCoNot (Brahman et al., 2024) | 379 |
| | PHTest (An et al., 2024) | 2,077 |
| | ORFuzzSet (Zhang et al., 2025a) | 1,788 |

*Table 1.* Evaluation datasets. Note that we sample 2,000 instances from the WildJailbreak test set.

### 2.1. Setup

**Notation.** Given a chat-templated prompt $\mathbf{x}=(x_1,\ldots,x_P)$, a language model autoregressively generates a completion $\mathbf{y} = (y_1,\ldots,y_T)$. For reasoning models, model-specific control tokens such as `<think>` and `</think>` delimit a thinking block followed by a final answer. After parsing these delimiters, we write the thinking trace as $\mathbf{r} = (r_1,\ldots,r_\tau)$ and the final response as $\mathbf{a}$, where $\tau$ denotes thinking compute; i.e., the number of tokens used in the thinking block. Refusal/compliance labels are assigned only to $\mathbf{a}$, not to the intermediate thinking trace $\mathbf{r}$.

For the thinking trace $\mathbf{r}$, we use $\mathbf{h}_t$ to denote the hidden representation of token $r_{1+\lfloor t(\tau-1)/99 \rfloor}$ in the final layer of the language model, where $t$ varies in $\{0,\ldots,99\}$. We will frequently use $\mathbf{h}_0$ to denote the final-layer hidden representation of the first token $r_1$ in the thinking trace; in practice, $r_1$ is typically a generic token such as "Okay."

**Evaluation.** We consider a suite of models including Qwen3-8B (Yang et al., 2025), OLMo-3-7B-Think (Team OLMo, 2025), Phi-4-Reasoning (Abdin et al., 2025), and GPT-OSS-20B (OpenAI, 2025a), chosen to span a range of model sizes and families. Our evaluation uses 2,500 harmful prompts for computing ASR and 2,885 benign prompts for computing ORR, with dataset composition shown in Table 1. For analyses requiring finer-grained estimates on benign prompts, we additionally use PHTest and ORFuzzSet, increasing the benign pool to 6,750 instances. Refusal/compliance labels are applied only to the model's final response. We compute a refusal score as the fraction of four guardrail classifiers that label the response as a refusal: WildGuard (Han et al., 2024), Qwen3Guard (Zhao et al., 2025a), Granite Guardian 3.3-8B (Padhi et al., 2024), and GPT-OSS-Safeguard-20B (OpenAI, 2025b;a). For instances where binary labels are needed (§2.2), an instance is marked

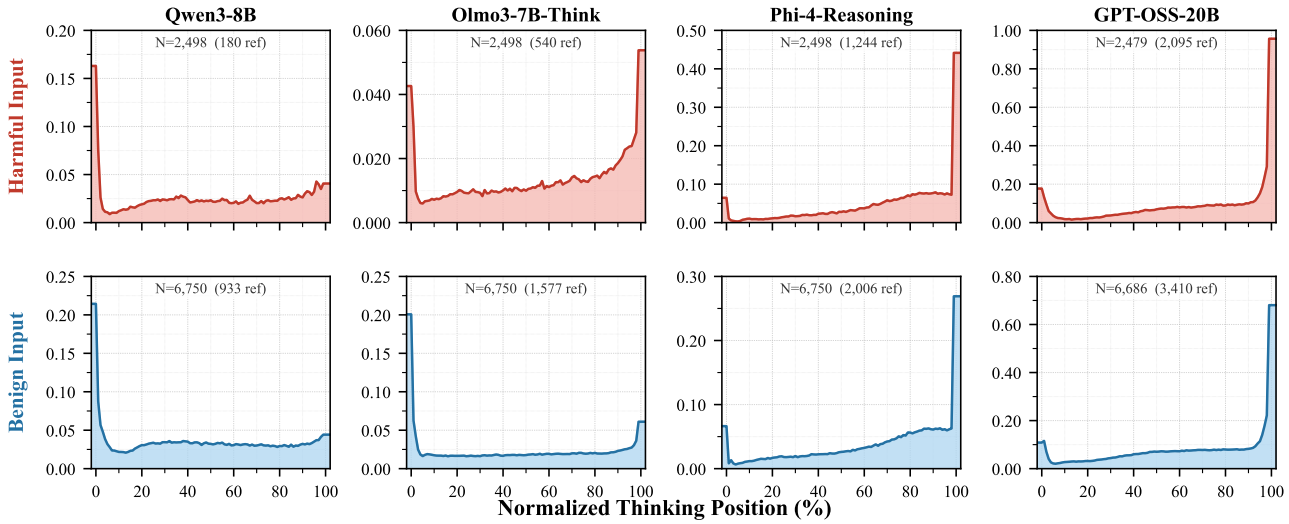


*Figure 1.* Fisher discriminant ratio $J(t)$ between eventual refusal and compliance traces, plotted against normalized position $t$ in the reasoning trace. Across all models, $J(t)$ traces the same U-shape (the refusal valley): high at the first generation position, dropping through the body of the reasoning trace, and recovering as thinking concludes.

a refusal if at least three of the four classifiers agree, with 2-2 ties resolved as compliance. Additional evaluation details are deferred to §B.1.[1]

### 2.2. Safety Decisions are Predictable at the First Token Representation

Our first question is: at what stage in a reasoning model's thinking does the refusal-compliance decision emerge in the model's representations? We examine this by probing the directional geometry of last-layer hidden states at each token on refusal and compliance thinking traces. For each model, we generate one completion per prompt in the full ASR pool and expanded ORR pool of §2.1. We label each completion's final response using the four-guardrail vote and partition thinking traces into refusal and compliance groups within each pool.

We consider two measures at each position $t$ between refusal and compliance groups. First, we use the multivariate Fisher discriminant to compute a whitened distance between the mean representations of the two groups:

$$J(t) \;=\; \big(\boldsymbol{\mu}_R(t) - \boldsymbol{\mu}_C(t)\big)^\top \boldsymbol{\Sigma}_W(t)^{-1} \big(\boldsymbol{\mu}_R(t) - \boldsymbol{\mu}_C(t)\big),$$

where $R$ and $C$ respectively denote the refusal and compliance groups, $\boldsymbol{\mu}_R(t), \boldsymbol{\mu}_C(t)$ are class-conditional means of $\mathbf{h}_t$, and $\boldsymbol{\Sigma}_W(t)$ is within-class covariance. More concretely,

$$\begin{aligned}
&\boldsymbol{\mu}_R(t) = \mathbb{E}_R \mathbf{h}_t; \quad \boldsymbol{\mu}_C(t) = \mathbb{E}_C \mathbf{h}_t; \\
&\boldsymbol{\Sigma}_W(t) = \mathbb{E}_C(\mathbf{h}_t - \boldsymbol{\mu}_C(t))(\mathbf{h}_t - \boldsymbol{\mu}_C(t))^\top \\
&\qquad\qquad + \mathbb{E}_R(\mathbf{h}_t - \boldsymbol{\mu}_R(t))(\mathbf{h}_t - \boldsymbol{\mu}_R(t))^\top.
\end{aligned}$$

Larger $J(t)$ indicates greater linear separability. Second, we train an $\ell_2$-regularized logistic regression probe on the first thinking-token representation $\mathbf{h}_0$ across both harmful and benign prompts to predict the guardrail-majority-vote's label on the model's final response. We evaluate this probe with 5-fold stratified cross-validation, yielding one out-of-fold refusal probability for every example. We report two metrics on these held-out probabilities: AUROC, a threshold-free measure of linear separability, and balanced accuracy (BAcc), a thresholded measure of probe accuracy. For BAcc, each held-out fold is binarized using a Youden-$J$ threshold[2] chosen from out-of-fold predictions on the other four folds, ensuring the threshold is selected without train-test contamination and accounts for class imbalance.

**Refusal/Compliance is Strongly Predictable at the First Token Representation.** In Figure 1, we observe that $J(t)$ shows a notable *valley* shape that begins high at $t$=0, drops through the body of the reasoning trace, and rises at $t$=99 across all models. The peak at the start suggests that the representation already exhibits a clear refusal/compliance outcome before any thinking tokens are generated. We include plots for additional models in Appendix C.1.

Strikingly, this early decision closely matches the refusal/-compliance outcome of the model's final response. As shown in Table 2, the probe on the first token representation

[1] Wherever possible, statistical uncertainty in our results will be reported using 95% bootstrap confidence intervals (CIs), derived via measuring the 2.5% and 97.5% percentiles from 1000 random resamples with replacement.

[2] See Appendix B for additional details.

| Model | Probe | Harmful | | Benign | | Pooled | |
|---|---|---|---|---|---|---|---|
| | | AUROC | BAcc | AUROC | BAcc | AUROC | BAcc |
| Qwen3-8B | Rep. | $0.915 \pm 0.020$ | $0.831 \pm 0.032$ | $0.961 \pm 0.008$ | $0.913 \pm 0.010$ | $0.953 \pm 0.007$ | $0.900 \pm 0.010$ |
| | Text | $0.492 \pm 0.042$ | $0.502 \pm 0.005$ | $0.502 \pm 0.019$ | $0.502 \pm 0.002$ | $0.499 \pm 0.017$ | $0.502 \pm 0.002$ |
| Olmo3-7B-Think | Rep. | $0.830 \pm 0.019$ | $0.742 \pm 0.023$ | $0.952 \pm 0.006$ | $0.877 \pm 0.010$ | $0.925 \pm 0.006$ | $0.842 \pm 0.009$ |
| | Text | $0.500 \pm 0.000$ | $0.500 \pm 0.000$ | $0.500 \pm 0.000$ | $0.500 \pm 0.000$ | $0.500 \pm 0.000$ | $0.500 \pm 0.000$ |
| Phi-4-Reasoning | Rep. | $0.838 \pm 0.015$ | $0.750 \pm 0.017$ | $0.855 \pm 0.010$ | $0.782 \pm 0.011$ | $0.856 \pm 0.008$ | $0.779 \pm 0.009$ |
| | Text | $0.561 \pm 0.022$ | $0.537 \pm 0.011$ | $0.502 \pm 0.014$ | $0.516 \pm 0.004$ | $0.515 \pm 0.013$ | $0.518 \pm 0.004$ |
| GPT-OSS-20B | Rep. | $0.838 \pm 0.021$ | $0.742 \pm 0.026$ | $0.901 \pm 0.008$ | $0.819 \pm 0.009$ | $0.899 \pm 0.006$ | $0.823 \pm 0.008$ |
| | Text | $0.678 \pm 0.031$ | $0.671 \pm 0.025$ | $0.619 \pm 0.013$ | $0.597 \pm 0.009$ | $0.601 \pm 0.012$ | $0.592 \pm 0.008$ |

*Table 2.* First-token probe behavior decomposed by probe type. We report AUROC and balanced accuracy (BAcc) values with 95% bootstrap CI (1,000 resamples). Rep. uses $\mathbf{h}_0$, and Text uses the first thinking-token TF-IDF feature. A linear probe on $\mathbf{h}_0$ predicts the final refusal/compliance decision, while text-only controls have little predictive power in either metric.

achieves high AUROC across all models ($0.856$–$0.953$), indicating that the two outcomes are linearly separable at $t{=}0$. Complementing this, the probe also achieves a balanced accuracy (BAcc) of $0.778$–$0.900$, showing that the first token's hidden state supports not only rank-order prediction but also an accurate binary readout of the final decision. As a surface-form control, we train an analogous probe on the TF-IDF representation of the first generated token. This text-only probe remains near chance in both AUROC and BAcc, ruling out the trivial explanation that the representational separability is a spurious artifact of the generated word.

### 2.3. Does Thinking Change the Safety Behavior?

The previous section shows that the hidden-state representation $\mathbf{h}_0$ already encodes the model's final decision well. At a high level, this predictability is not by itself a failure mode of reasoning models, as certain prompts may be unambiguous enough that the correct decision can be made trivially. However, in general, the thinking process should still be able to affect the model's behavior rather than merely rationalize an initial decision. Thus, we investigate whether the model's thinking process serves to meaningfully revise this initial decision.

We first test how the safety behavior of the model changes with the amount of thinking compute by examining behavior on rollouts at different thinking budgets on 200 harmful and benign prompts per model. For each prompt, we sample $m = 8$ independent thinking traces, truncate each trace at a budget $B \in \{20\%, 40\%, 60\%, 80\%\}$ of the original trace length. For each truncated thinking prefix, we sample $n{=}8$ final responses conditioned on the prefix and label the responses accordingly with guardrails. We compute ASR and ORR for each budget $B$ by first averaging refusal/compliance labels over continuations sampled from a fixed prefix, then averaging these estimates over prefixes at the corresponding budget rate.

**Increasing Thinking Budget Doesn't Necessarily Improve Safety Behavior.** We show results in Table 3, which shows that increasing thinking budget does not produce a consistent improvement in safety behavior. Qwen3-8B and OLMo-3-7B-Think are essentially flat as the thinking budget increases, while Phi-4-Reasoning exhibits an over-compliance behavior where longer thinking reduces over-refusal (ORR) on benign prompts but increases attack success (ASR) on harmful prompts. GPT-OSS-20B is the only partial exception where ASR is reduced with longer thinking, but with ORR remaining largely unchanged. The current evidence suggests that increasing thinking compute does not reliably make models both safer on harmful prompts and more helpful on benign prompts.

**Thinking Does Not Change the Final Decision.** The above also allows us to examine the sources of variance in the model's decisions. A high variance in response labels conditioned on a given prefix suggests that the model continues to deliberate on its decision as the thinking trace unfolds.

For a fixed prompt and budget $B$, let $z_{m,B,k} \in \{0, 1\}$ denote whether the $k$-th final response from sampled thinking prefix $m$ with budget $B$ is labeled as a refusal, and let $q_{m,B} = \frac{1}{n}\sum_k z_{m,B,k}$ be the refusal rate after conditioning on that prefix. We measure *within-prefix variance*, which shows whether different continuations from the same prefix lead to different outcomes:

$$\mathrm{WithinPrefix}(B) = \mathbb{E}_m[4q_{m,B}(1 - q_{m,B})]\,,$$

where the factor of $4$ normalizes Bernoulli variance so that the maximum possible variance is $1$.

| Model | Pool | 20% | 40% | 60% | 80% | Δ |
|---|---|---|---|---|---|---|
| Qwen3-8B | ASR ↓ | $90.7 \pm 2.7$ | $90.8 \pm 2.5$ | $90.2 \pm 2.7$ | $89.1 \pm 2.6$ | $-1.6 \pm 1.0$ |
| | ORR ↓ | $3.3 \pm 1.3$ | $3.2 \pm 1.4$ | $3.8 \pm 1.5$ | $5.4 \pm 1.8$ | $+2.1 \pm 0.9$ |
| OLMo-3-7B-Think | ASR ↓ | $86.8 \pm 3.1$ | $86.0 \pm 3.4$ | $86.3 \pm 3.2$ | $86.4 \pm 3.1$ | $-0.3 \pm 1.3$ |
| | ORR ↓ | $8.7 \pm 2.5$ | $8.5 \pm 2.5$ | $8.6 \pm 2.6$ | $8.5 \pm 2.2$ | $-0.2 \pm 1.1$ |
| Phi-4-Reasoning | ASR ↓ | $60.9 \pm 5.4$ | $62.8 \pm 5.4$ | $64.3 \pm 5.0$ | $64.6 \pm 5.5$ | $+4.1 \pm 1.8$ |
| | ORR ↓ | $35.7 \pm 4.4$ | $29.4 \pm 4.6$ | $25.3 \pm 4.6$ | $23.7 \pm 4.2$ | $-12.0 \pm 2.6$ |
| GPT-OSS-20B | ASR ↓ | $62.3 \pm 5.2$ | $55.2 \pm 5.4$ | $50.7 \pm 5.3$ | $46.9 \pm 5.2$ | $-14.8 \pm 3.1$ |
| | ORR ↓ | $33.2 \pm 4.7$ | $33.0 \pm 4.7$ | $32.7 \pm 4.6$ | $32.5 \pm 4.6$ | $-1.0 \pm 2.3$ |

*Table 3.* Effect of relative thinking budget on final safety behavior. We truncate sampled thinking traces at 20–80% of their natural length, sample continuations, and report ASR on harmful prompts and ORR on benign prompts under guardrail-majority labeling. Values are percentages with 95% bootstrap CIs; Δ denotes the change from 20% to 80%. Lower is better for both ASR and ORR.

Our results are included in Figure 2. We observe that within-prefix variance is already near zero at $B{=}20\%$ ($<0.1$) and decreases further as the prefix length grows across all models. Independently sampled continuations generally result in the same outcome after conditioning on even a short reasoning prefix, suggesting that later thinking rarely changes the model's final response decision.

### 2.4. Sentence-Level Analysis of Thinking Traces

§2.3 suggests that the thinking trace primarily acts as a post-hoc rationalization of an early decision, rather than a mechanism that guides the model toward a correct final response. This raises a natural question: what function, then, do the individual sentences within the thinking trace serve?

To quantify this, we analyze safety deliberation at the sentence level within thinking traces. We segment each trace into sentences $c_{1:T}$ and use GPT-5.4 to assign each sentence a stance label in $\{-1, 0, +1\}$, corresponding to refusal-leaning, neutral, and compliance-leaning text, respectively. We first measure the fraction of traces that exhibit at least one *oscillation*, i.e., a change in stance from refusal-leaning to compliance-leaning or vice versa.

On instances where we find an oscillation, we now measure whether these oscillations have a significant effect on the final response of the model. For each oscillation, let $i_{\text{pre}}$ denote the index of the *last* sentence with the previous stance, and let $i_{\text{post}}$ denote the index of the *last* sentence with the new stance. We define the corresponding prefixes of the thinking trace as $x^{\text{pre}} = c_{1:i_{\text{pre}}}$ and $x^{\text{post}} = c_{1:i_{\text{post}}}$. Following §2.3, we estimate the refusal rates $q_{x^{\text{pre}}}$ and $q_{x^{\text{post}}}$ using $K{=}32$ final-response generations. We quantify the effect of the segment between $i_{\text{pre}}$ and $i_{\text{post}}$ using its *stance-aligned shift*:

$$\Delta\hat{q}_{i_{\text{pre}}:i_{\text{post}}} = s_{\text{post}} \left( q_{x^{\text{pre}}} - q_{x^{\text{post}}} \right),$$

where $s_{\text{post}} = +1$ for compliance→refusal transitions and $s_{\text{post}} = -1$ for refusal→compliance transitions. By construction, $\Delta\hat{q} \in [-1, 1]$, and positive values indicate that the intervening segment shifts the final response distribution in the direction implied by the new stance. For each trace, we collect all $\Delta\hat{q}$ values and evaluate their statistical significance using $p$-values (Appendix B). We then report the fraction of traces containing at least one segment that produces a statistically significant shift in the model's refusal decision.

**Most Thinking Trace Segments Show No Significant Effect.** We present the results in Figure 3. Observe that GPT-OSS-20B and Phi-4-Reasoning exhibit very little oscillation in their thinking traces; in fact, more than 74% of traces for these models contain no stance changes at all. In contrast, Qwen3-8B and OLMo-3-7B-Think exhibit substantially more oscillatory behavior. However, even for these models, the vast majority of traces do not contain any segment that significantly affects the model's refusal probability. Overall, across all models, at most 15% of traces contain any segment whose effect on the final response is statistically significant.

## 3. What Do Existing Defenses Do?

The results in §2 suggest that safety decisions in LRMs are often largely determined at the first token's representation, and that additional thinking does not reliably move the final decision toward the desired regime of both lower ASR and ORR. Now, we ask how existing safety defenses interact with this behavior. An effective defense should enable the thinking trace to be truly utilized for safety deliberation and revise or refine its initial safety judgment instead of merely rationalizing it. We examine whether this occurs for two classes of defenses used in practice: inference-time interventions and training-based defenses based on supervised fine-tuning (SFT) or reinforcement learning with human feedback (RLHF).

### 3.1. Inference-Based Defenses

Inference-time defenses provide a direct test of whether LRM safety can be improved by intervening on the thinking trace itself. Most existing defenses of this form modify the

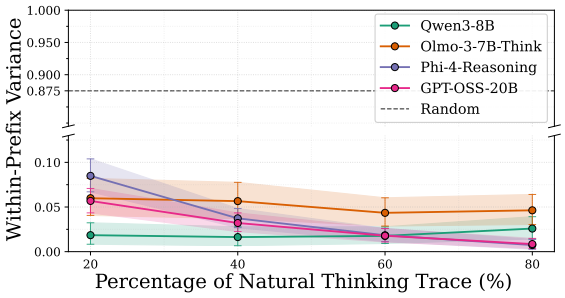


*Figure 2.* Within-prefix variance with 95% bootstrap CIs; We pool over harmful and benign instances. Dashed line indicates fair-coin reference 0.875 ($n = 8$).

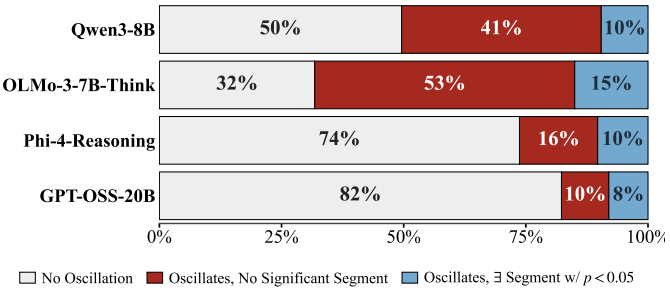


*Figure 3.* Per-trace classification of audited stance traces. Across all four models, only 8–15% of traces contain segments that have a statistically significant effect on the model's refusal rate.

thinking trace by inserting safety-relevant text at selected positions during generation: **SafePath ZS** (Jeung et al., 2025) prepends a short safety primer at the start of the reasoning block; **PSR** (Phan et al., 2025) interleaves generation with periodic self-reflection checkpoints; and **SafeRemind** (Kim et al., 2026) dynamically injects safe-reminding phrases during thinking when entropy indicates a decision-locking point. The common premise is that making safety considerations salient within the trace should steer subsequent reasoning toward safer final behavior. See Appendix B.2 for further details on implementation.

However, §2.3–§2.4 suggests a more limited result. If LRMs do not use the trace to deliberate on the safety decision, we would expect such interventions may increase refusal without improving the ASR–ORR trade-off. Table 4 is consistent with this prediction: inference-time defenses either lower ASR at the cost of higher ORR, or have limited effect on the final response of the model. The tradeoff is most pronounced for OLMo-3-7B-Think; for the remaining models, the effects are weaker but similarly trade harmful-prompt safety against benign-prompt helpfulness. For other models, the effect is more limited, and although reasoning shifts toward increased refusal of harmful prompts, this also comes at the cost of decreased helpfulness on benign prompts.

### 3.2. Training-Based Defenses

Training-based defenses are a more direct approach to modifying the model's safety behavior than inference-time interventions. We consider defenses mainly from two classes of methods. The first are supervised fine-tuning (SFT) methods that center around training the model on well-curated gold reasoning traces:

- **STAR-1** (Wang et al., 2025b) constructs a 1K-example policy-grounded safety-reasoning dataset.
- **R1-ACT** (In et al., 2025) fine-tunes on traces that explicitly insert a harmfulness-assessment step to activate latent safety knowledge.
- **SafeKey** (Zhou et al., 2025b) augments safety-trace SFT with auxiliary objectives targeting the emergence of a "safety aha" key sentence.
- **ThinkSafe** (Lee et al., 2026) uses refusal steering to elicit in-distribution safety traces from the model itself before fine-tuning on those self-generated responses.

The second are reinforcement learning with human feedback (RLHF)-style or preference-optimization methods that further optimize the model using preference or process-level safety signals:

- **STAIR** (Zhang et al., 2025b) combines structured safety reasoning with iterative preference optimization and process-reward-guided search.
- **RAPO** (Wei et al., 2026) constructs preference pairs that favor reasoning traces whose amount and granularity of safety analysis match the prompt's risk level, and optimizes the model to produce more risk-appropriate safety reasoning.

We include further details for each method in Appendix B.3.

**Training Does Not Reliably Improve Safety Discrimination.** Table 4 shows that existing defenses primarily shift the trade-off between ASR and ORR rather than improving it. The direction and magnitude of this shift are strongly model-dependent, but the underlying trade-off remains consistent: gains along one axis typically come at the expense of the other. On Qwen3-8B and OLMo-3-7B-Think, several training-based methods substantially reduce harmful compliance, but with large increases in over-refusal (e.g., ThinkSafe on Qwen3-8B: ASR 84.8 → 30.8, ORR 7.6 → 45.5; RAPO on OLMo-3-7B-Think: ASR 70.6 → 17.7, ORR 21.2 → 49.9). Conversely, some methods shift models in the opposite direction, reducing over-refusal while increasing harmful compliance (e.g., STAR1 on GPT-OSS-20B: ASR 26.5 → 71.7, ORR 44.9 → 14.8). These results suggest that current defenses do not improve the refusal-compliance trade-off, but instead shift it, high-

| | Method | Qwen3-8B ASR ↓ | Qwen3-8B ORR ↓ | OLMo-3-7B-Think ASR ↓ | OLMo-3-7B-Think ORR ↓ | Phi-4-Reasoning ASR ↓ | Phi-4-Reasoning ORR ↓ | GPT-OSS-20B ASR ↓ | GPT-OSS-20B ORR ↓ |
|---|---|---|---|---|---|---|---|---|---|
| | Base | 84.8 | 7.6 | 70.6 | 21.2 | 37.2 | 49.0 | 26.5 | 44.9 |
| Inf. | PSR | 86.90 ↑2.1 | 6.22 ↓1.3 | 37.72 ↓32.9 | 44.19 ↑23.0 | 49.99 ↑12.8 | 49.17 ↑0.1 | 26.16 ↓0.3 | 51.03 ↑6.2 |
| Inf. | SafeRemind | 80.22 ↓4.6 | 11.16 ↑3.6 | 33.11 ↓37.5 | 46.02 ↑24.8 | 54.25 ↑17.1 | 46.95 ↓2.1 | 18.83 ↓7.6 | 55.10 ↑10.2 |
| Inf. | SafePath-ZS | 68.51 ↓16.3 | 13.41 ↑5.9 | 30.34 ↓40.3 | 49.34 ↑28.2 | 46.14 ↑9.0 | 52.95 ↑3.9 | 14.92 ↓11.5 | 59.04 ↑14.2 |
| Training | STAR1 | 41.0 ↓43.8 | 45.5 ↑37.9 | 39.2 ↓31.4 | 49.8 ↑28.6 | 35.8 ↓1.4 | 45.3 ↓3.7 | 71.7 ↑45.2 | 14.8 ↓30.1 |
| Training | SafeKey | 56.2 ↓28.6 | 29.1 ↑21.5 | 48.1 ↓22.5 | 36.9 ↑15.7 | 52.7 ↑15.5 | 33.5 ↓15.5 | 71.0 ↑44.5 | 19.4 ↓25.5 |
| Training | R1ACT | 57.0 ↓27.8 | 37.8 ↑30.3 | 85.4 ↑14.8 | 27.7 ↑6.6 | 53.5 ↑16.4 | 39.2 ↓9.9 | 54.6 ↑28.1 | 46.7 ↑1.8 |
| Training | ThinkSafe | 30.8 ↓54.0 | 45.5 ↑37.9 | 46.7 ↓23.9 | 38.4 ↑17.2 | 39.6 ↑2.4 | 48.9 ↓0.2 | 18.8 ↓7.6 | 52.3 ↑7.5 |
| Training | RAPO | 34.7 ↓50.1 | 35.0 ↑27.5 | 17.7 ↓52.9 | 49.9 ↑28.7 | 56.1 ↑18.9 | 51.3 ↑2.3 | 33.8 ↑7.3 | 47.6 ↑2.7 |
| Training | STAIR | 53.8 ↓31.0 | 45.3 ↑37.8 | 54.6 ↓16.0 | 38.3 ↑17.1 | 37.1 ↓0.1 | 48.7 ↓0.4 | 26.4 ↓0.1 | 68.6 ↑23.7 |

*Table 4.* ASR and ORR of inference-time and training-time defenses under 4-guardrail fractional-vote labeling. The Base row uses the inference-time base model results. ASR is compliance on harmful prompts; ORR is refusal on benign prompts. Lower is better for both metrics. Gray arrows and cell colors indicate change relative to Base: green denotes decreases and red denotes increases, with darker shading indicating larger absolute changes.

lighting the need for methods that can shift the tradeoff frontier outward.

**Training-Based Defenses Increase Refusal-Leaning Traces.** To better understand the behavioral effects of these defenses, we examine how the generated thinking traces change after training. We apply the same annotation procedure from §2.4, using GPT-5.4 to label each thinking-trace sentence as neutral, refusal-leaning, or compliance-leaning. For each model-defense pair, we compute fractions of each stance on each thinking trace separately on harmful and benign prompts. Averaging over traces yields three fractions, $(f_{\text{refuse}}, f_{\text{neutral}}, f_{\text{comply}})$, with $f_{\text{refuse}} + f_{\text{neutral}} + f_{\text{comply}} = 1$ by construction. We compare each stance fraction to the corresponding behavioral refusal rate, using $1-\text{ASR}$ for harmful prompts and ORR for benign prompts, and report Pearson's correlation (r).

Results are shown in Figure 4. Most defense methods increase the ratio of refusal-leaning sentences along with refusal rates, while largely preserving the fraction of neutral sentences and decreasing the fraction of compliance-leaning sentences. Observe that the fraction of refusal-leaning trace text is strongly positively correlated with refusal behavior on both harmful and benign prompts ($r = +0.72$ and $r = +0.54$, respectively). Conversely, compliance-leaning text is negatively correlated with refusal behavior ($r = -0.75$ and $r = -0.50$), while neutral stance is only weakly anti-correlated ($r = -0.36$ and $r = -0.32$). This pattern suggests that many training-based defenses may be improving harmful prompt safety simply by teaching the model to generate more refusal-leaning traces, rather than by making the trace a more deliberative procedure.

## 4. Discussion

We provide evidence that current LRMs do not effectively use thinking for safety deliberation, and that existing defenses may not induce such deliberation reliably. We show that the hidden representation of the first thinking token already strongly predicts a refusal/compliance decision, and that this prediction closely matches the model's final response. Additional thinking does not consistently improve safety decisions, and seemingly deliberative segments exert limited influence on the final outcome. Moreover, many existing defenses primarily steer models toward more refusal-leaning behavior, often increasing over-refusal on benign prompts. Together, these findings suggest that future defenses should explicitly train models to use thinking traces for meaningful safety deliberation rather than merely rationalize an early-formed decision. Below, we discuss related work, limitations, and future directions.

**Related Works.** More broadly, our work contributes to the growing literature on the safety properties of large reasoning models and reasoning-based alignment strategies (Wang et al., 2025a; Zhou et al., 2025a; Guan et al., 2024). Our findings connect to several recent lines of work on the safety and interpretability of instruction-tuned and reasoning language models, which we discuss below. We further defer additional discussions to Appendix §A.

First, several prior works have shown that safety alignment in instruction-tuned language models can be surprisingly shallow, with refusal behavior often determined by a small number of early tokens and vulnerable to simple prefix manipulations (Qi et al., 2025; Liu et al., 2026; Zhao et al., 2025c; Yin et al., 2025). Our work extends these observations to reasoning models. In particular, we show that refusal/compliance behavior is already strongly encoded at the beginning of the thinking trace, and current reasoning models make only limited use of their thinking capabilities for safety deliberation.

Second, our work is closely related to studies on the faith-

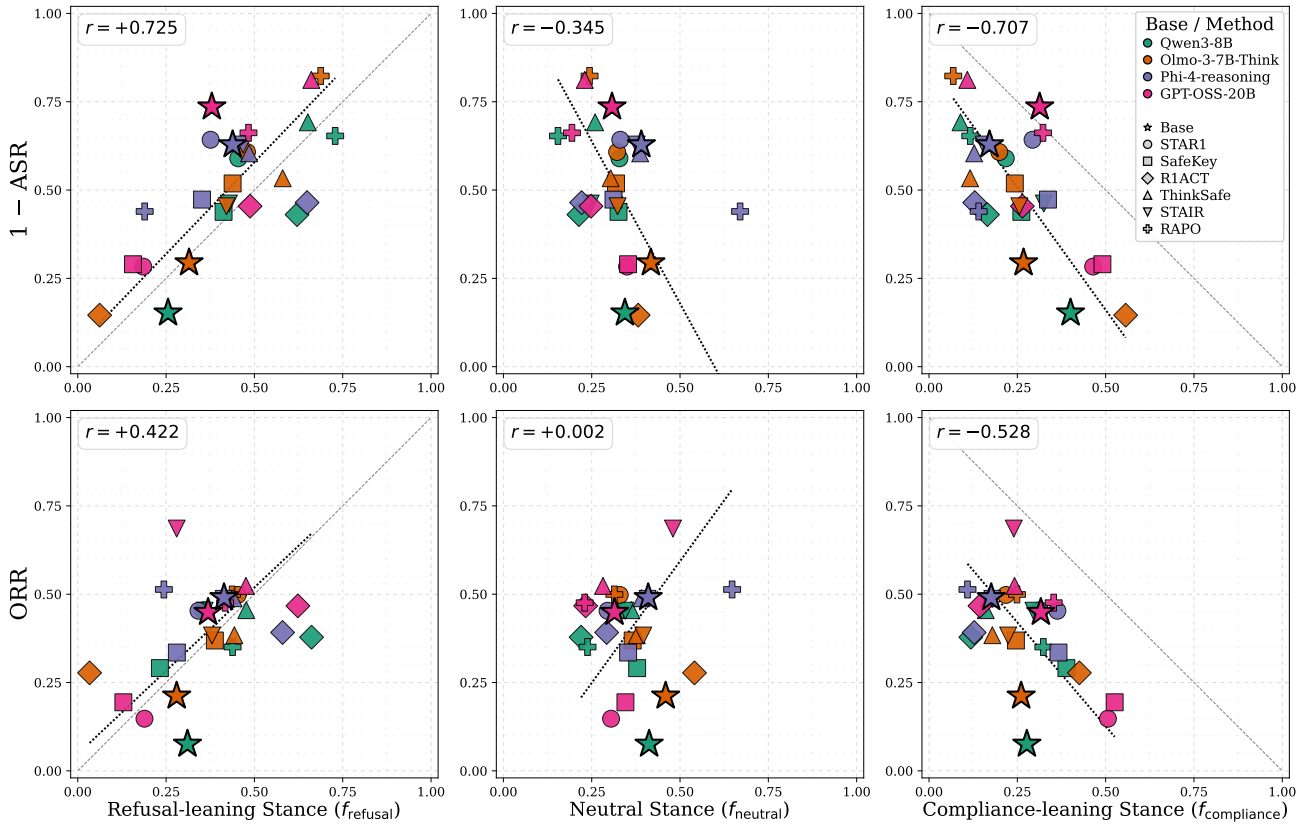


*Figure 4.* Trace stance predicts behavioral refusal across training-based defenses. Each scatter point corresponds to a model-method pair. $x$-axis measures the fraction of sentence-level thinking segments labeled for each stance, and $y$-axis measures the final refusal ($1 - \text{ASR}$) and compliance rates (ORR). Dashed lines show reduced major-axis fits for visualization only; reported associations are Pearson correlations.

fulness of chain-of-thought reasoning. In reasoning tasks, (Boppana et al., 2026) and (Cox et al., 2026) show that, for easier questions, the final answer can often be decoded from hidden representations of the initial tokens, suggesting that later reasoning tokens may partially rationalize an already-determined decision. However, these effects become substantially weaker on harder questions. Our work provides complementary evidence in the context of safety: although reasoning traces may contain apparent safety deliberation, these intermediate deliberative segments exert only limited influence on the final response.

**Limitations.** Our study has two limitations. First, we exclusively focus on open-weight LRMs of moderate sizes, which calls into question whether significantly larger reasoning models exhibit the same behavior. However, prior work suggests that scale alone should not be assumed to resolve safety failures as they find that larger models can remain unreliable under safety-relevant conditions, and that refusal robustness under adversarial pressure is not guaranteed by model size (Zhou et al., 2024; Anil et al., 2024; Wang, 2026). Second, our evaluation is also limited to refusal/compliance behavior on harmful and benign prompt sets. This captures the central ASR–ORR trade-off, but does not cover all dimensions of safety, such as factuality, deception, privacy, or multi-turn behavior.

**Future Work.** While §2 shows that refusal-or-compliance behavior is strongly predictable early in generation, it is not immediately clear where in the training process this originates. Understanding whether it arises during pretraining, instruction tuning, or post-training is an important avenue for investigation. From §3, we observe that simply supervising on gold traces, mixing in both harmful and benign data, or optimizing with respect to final labels may be insufficient to induce true safety deliberation. Hence, a question of interest that arises is how to design training objectives to reward true deliberation. We leave both as promising directions for future work.

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

## References


Abdin, M., Agarwal, S., Awadallah, A., Balachandran, V., Behl, H., Chen, L., de Rosa, G., Gunasekar, S., Javaheripi, M., Joshi, N., et al. Phi-4-reasoning technical report, 2025.

An, B., Zhu, S., Zhang, R., Panaitescu-Liess, M.-A., Xu, Y., and Huang, F. Automatic pseudo-harmful prompt generation for evaluating false refusals in large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=ljFgX6A8NL.

Anil, C., Durmus, E., Panickssery, N., Sharma, M., Benton, J., Kundu, S., Batson, J., Tong, M., Mu, J., Ford, D., et al. Many-shot jailbreaking. *Advances in Neural Information Processing Systems*, 37:129696–129742, 2024.

Arcuschin, I., Janiak, J., Krzyzanowski, R., Rajamanoharan, S., Nanda, N., and Conmy, A. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*, 2025. URL https://arxiv.org/abs/2503.08679.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., El Showk, S., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

Baker, B., Huizinga, J., Gao, L., Dou, Z., Guan, M. Y., Madry, A., Zaremba, W., Pachocki, J., and Farhi, D. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025. URL https://arxiv.org/abs/2503.11926.

Boppana, S., Ma, A., Loeffler, M., Sarfati, R., Bigelow, E., Geiger, A., Lewis, O., and Merullo, J. Reasoning theater: Disentangling model beliefs from chain-of-thought, 2026. URL https://arxiv.org/abs/2603.05488.

Brahman, F., Kumar, S., Balachandran, V., Dasigi, P., Pyatkin, V., Ravichander, A., Wiegreffe, S., Dziri, N., Chandu, K., Hessel, J., et al. The art of saying no: Contextual noncompliance in language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2407.12043.

Cox, K., Kianersi, D., and Garriga-Alonso, A. Decoding answers before chain-of-thought: Evidence from pre-cot probes and activation steering, 2026. URL https://arxiv.org/abs/2603.01437.

Cui, G., Yuan, L., Ding, N., Yao, G., He, B., Zhu, W., Ni, Y., Xie, G., Xie, R., Lin, Y., Liu, Z., and Sun, M. Ultrafeedback: Boosting language models with scaled ai feedback, 2023. URL https://arxiv.org/abs/2310.01377.

Cui, J., Chiang, W.-L., Stoica, I., and Hsieh, C.-J. OR-Bench: An over-refusal benchmark for large language models. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2405.20947.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Emmons, S., Jenner, E., Elson, D. K., Saurous, R. A., Rajamanoharan, S., Chen, H., Shafkat, I., and Shah, R. When chain of thought is necessary, language models struggle to evade monitors. *arXiv preprint arXiv:2507.05246*, 2025. URL https://arxiv.org/abs/2507.05246.

Guan, M. Y., Joglekar, M., Wallace, E., Jain, S., Barak, B., Helyar, A., Dias, R., Vallone, A., Ren, H., Wei, J., Chung, H. W., Toyer, S., Heidecke, J., Beutel, A., and Glaese, A. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*, 2024.

Han, S., Rao, K., Ettinger, A., Jiang, L., Lin, B. Y., Lambert, N., Choi, Y., and Dziri, N. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0261.

Huan, M., Li, Y., Zheng, T., Xu, X., Kim, S., Du, M., Poovendran, R., Neubig, G., and Yue, X. Does math reasoning improve general LLM capabilities? understanding transferability of LLM reasoning. *arXiv preprint arXiv:2507.00432*, 2025.

Hughes, J., Price, S., Lynch, A., Schaeffer, R., Barez, F., Koyejo, S., Sleight, H., Jones, E., Perez, E., and Sharma, M. Best-of-n jailbreaking. *arXiv preprint arXiv:2412.03556*, 2024.

In, Y., Kim, W., Park, S., and Park, C. R1-ACT: Efficient reasoning model safety alignment by activating safety knowledge, 2025. URL https://arxiv.org/abs/2508.00324.

Jeung, W., Yoon, S., Kahng, M., and No, A. SAFEPATH: Preventing harmful reasoning in chain-of-thought via early alignment, 2025. URL https://arxiv.org/abs/2505.14667.

Ji, J., Hong, D., Zhang, B., Chen, B., Dai, J., Zheng, B., Qiu, T., Zhou, J., Wang, K., Li, B., Han, S., Guo, Y., and Yang, Y. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference, 2024. URL https://arxiv.org/abs/2406.15513.

Jiang, F., Xu, Z., Li, Y., Niu, L., Xiang, Z., Li, B., Lin, B. Y., and Poovendran, R. Safechain: Safety of language models with long chain-of-thought reasoning capabilities, 2025. URL https://arxiv.org/abs/2502.12025.

Jiang, L., Rao, K., Han, S., Ettinger, A., Brahman, F., Kumar, S., Mireshghallah, N., Lu, X., Sap, M., Choi, Y., and Dziri, N. WildTeaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.18510.

Kim, S.-H., Jin, H., Lee, Y., and Han, Y.-S. How does the thinking step influence model safety? an entropy-based safety reminder for LRMs, 2026. URL https://arxiv.org/abs/2601.03662.

Kim, T., Tajwar, F., Raghunathan, A., and Kumar, A. Reasoning as an adaptive defense for safety. In *NeurIPS 2025 Workshop: Reliable ML from Unreliable Data*, 2025. URL https://openreview.net/forum?id=fbqpswqlI5.

Knight, C. Q., Deshpande, K., Sirdeshmukh, V., Mankikar, M., Team, S. R., Team, S. R., and Michael, J. Fortress: Frontier risk evaluation for national security and public safety, 2025. URL https://arxiv.org/abs/2506.14922.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2205.11916.

Korbak, T., Balesni, M., Barnes, E., Bengio, Y., et al. Chain of thought monitorability: A new and fragile opportunity for ai safety, 2025. URL https://arxiv.org/abs/2507.11473.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention, 2023. URL https://arxiv.org/abs/2309.06180.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. URL https://arxiv.org/abs/2307.13702.

Lee, S., Park, S., Choi, Y., Kim, G., Kang, M., Yun, J., Park, D., Park, J., and Hwang, S. J. THINKSAFE: Self-generated safety alignment for reasoning models, 2026. URL https://arxiv.org/abs/2601.23143.

Li, N., Han, Z., Steneker, I., Primack, W., Goodside, R., Zhang, H., Wang, Z., Menghini, C., and Yue, S. Llm defenses are not robust to multi-turn human jailbreaks yet. *arXiv preprint arXiv:2408.15221*, 2024.

Liu, S., Pei, H., and Liu, Z. Shallowjail: Steering jailbreaks against large language models, 2026. URL https://arxiv.org/abs/2602.07107.

Liu, X., Xu, N., Chen, M., and Xiao, C. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*, 2023.

Liu, Y., He, X., Xiong, M., Fu, J., Deng, S., and Hooi, B. Flipattack: Jailbreak llms via flipping. *arXiv preprint arXiv:2410.02832*, 2024.

Mehrotra, A., Zampetakis, M., Kassianik, P., Nelson, B., Anderson, H., Singer, Y., and Karbasi, A. Tree of attacks: Jailbreaking black-box llms automatically. *Advances in Neural Information Processing Systems*, 37:61065–61105, 2024.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025a. URL https://openai.com/index/gpt-oss-model-card/.

OpenAI. Technical report: Performance and baseline evaluations of gpt-oss-safeguard-120b and gpt-oss-safeguard-20b. Technical report, OpenAI, October 2025b. URL https://openai.com/index/gpt-oss-safeguard-technical-report/.

Padhi, I., Nagireddy, M., Cornacchia, G., Chaudhury, S., Pedapati, T., Dognin, P., Murugesan, K., Miehling, E., Santillán Cooper, M., Fraser, K., Zizzo, G., Hameed, M. Z., Purcell, M., Desmond, M., Pan, Q., Ashktorab, Z.,

Vejsbjerg, I., Daly, E. M., Hind, M., Geyer, W., Rawat, A., Varshney, K. R., and Sattigeri, P. Granite guardian, 2024.

Perez, F. and Ribeiro, I. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*, 2022.

Phan, H., Li, V., and Lei, Q. Think twice, generate once: Safeguarding by progressive self-reflection, 2025. URL https://arxiv.org/abs/2510.01270.

Qi, X., Panda, A., Lyu, K., Ma, X., Roy, S., Beirami, A., Mittal, P., and Henderson, P. Safety alignment should be made more than just a few tokens deep. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6Mxhg9PtDE.

Ren, Q., He, Q., Zhang, B., Zeng, J., Liang, J., Xiao, Y., Zhou, W., Sun, Z., and Yu, F. Beyond the trade-off: Self-supervised reinforcement learning for reasoning models' instruction following, 2025. URL https://arxiv.org/abs/2508.02150.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Tang, Y., Wang, S., Madaan, L., and Munos, R. Beyond verifiable rewards: Scaling reinforcement learning in language models to unverifiable data. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=pc6M9h3T9m.

Team OLMo. Olmo 3, 2025.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *NeurIPS*, 2023. URL https://arxiv.org/abs/2305.04388.

Wang, C., Liu, Y., Bi, B., Zhang, D., Li, Z.-Z., Ma, Y., He, Y., Yu, S., Li, X., Fang, J., Zhang, J., and Hooi, B. Safety in large reasoning models: A survey, 2025a. URL https://arxiv.org/abs/2504.17704.

Wang, Y. Scaling laws of refusal robustness: Why bigger LMs are not necessarily safer, 2026. URL https://openreview.net/forum?id=6x4DWjhMsH. ICLR 2026 desk-rejected submission, OpenReview.

Wang, Z., Tu, H., Wang, Y., Wu, J., Liu, Y., Mei, J., Bartoldson, B. R., Kailkhura, B., and Xie, C. STAR-1: Safer alignment of reasoning LLMs with 1k data, 2025b. URL https://arxiv.org/abs/2504.01903.

Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: How does llm safety training fail? *Advances in neural information processing systems*, 36:80079–80110, 2023.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2201.11903.

Wei, Z., Zhang, Q., Hu, X., and Xu, X. RAPO: Risk-aware preference optimization for generalizable safe reasoning, 2026. URL https://arxiv.org/abs/2602.04224.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report, 2025.

Yin, Q., Leong, C. T., Yang, L., Huang, W., Li, W., Wang, X., Yoon, J., YunXing, XingYu, and Gu, J. Refusal falls off a cliff: How safety alignment fails in reasoning?, 2025. URL https://arxiv.org/abs/2510.06036.

Yu, J., Lin, X., Yu, Z., and Xing, X. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*, 2023.

Yu, P., Lanchantin, J., Wang, T., Yuan, W., Golovneva, O., Kulikov, I., Sukhbaatar, S., Weston, J., and Xu, J. Cot-self-instruct: Building high-quality synthetic prompts for reasoning and non-reasoning tasks, 2025. URL https://arxiv.org/abs/2507.23751.

Zhang, H., Wang, D., Liu, Y., Chen, K., Wang, J., Ying, X., Liu, L., and Wang, W. Orfuzz: Fuzzing the "other side" of llm safety – testing over-refusal, 2025a. URL https://arxiv.org/abs/2508.11222.

Zhang, Y., Zhang, S., Huang, Y., Xia, Z., Fang, Z., Yang, X., Duan, R., Yan, D., Dong, Y., and Zhu, J. STAIR: Improving safety alignment with introspective reasoning. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=aHzPGyUhZa.

Zhang, Z., Xu, W., Wu, F., and Reddy, C. K. FalseReject: A resource for improving contextual safety and mitigating over-refusals in LLMs via structured reasoning. *arXiv preprint arXiv:2505.08054*, 2025c.

Zhao, H., Yuan, C., Huang, F., Hu, X., Zhang, Y., Yang, A., Yu, B., Liu, D., Zhou, J., Lin, J., et al. Qwen3guard technical report, 2025a.

Zhao, S., Duan, R., Liu, J., Jia, X., Wang, F., Wei, C., Cheng, R., Xie, Y., Liu, C., Guo, Q., Tao, J., Xue, H.,

and Wei, X. Strata-sword: A hierarchical safety evaluation towards llms based on reasoning complexity of jailbreak instructions, 2025b. URL https://arxiv.org/abs/2509.01444.

Zhao, X., Yang, X., Pang, T., Du, C., Li, L., Wang, Y.-X., and Wang, W. Y. Weak-to-strong jailbreaking on large language models, 2025c. URL https://arxiv.org/abs/2401.17256.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q. V., and Chi, E. H. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WZH7099tgfM.

Zhou, K., Liu, C., Zhao, X., Jangam, S., Srinivasa, J., Liu, G., Song, D., and Wang, X. E. The hidden risks of large reasoning models: A safety assessment of r1, 2025a. URL https://arxiv.org/abs/2502.12659.

Zhou, K., Zhao, X., Liu, G., Srinivasa, J., Feng, A., Song, D., and Wang, X. E. SafeKey: Amplifying aha-moment insights for safety reasoning, 2025b. URL https://arxiv.org/abs/2505.16186.

Zhou, L., Schellaert, W., Martínez-Plumed, F., Moros-Daval, Y., Ferri, C., and Hernández-Orallo, J. Larger and more instructable language models become less reliable. *Nature*, 634:61–68, 2024. doi: 10.1038/s41586-024-07930-y. URL https://www.nature.com/articles/s41586-024-07930-y.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

# A. Related Works

In this section, we provide extensive discussion on works that are closely connected to our work.

**Reasoning and Deliberative Alignment.** Recent advances in large reasoning models (LRMs) have shown that allocating inference-time compute through chain-of-thought or thinking traces can substantially improve performance on difficult reasoning tasks (Wei et al., 2022; Kojima et al., 2022; Zhou et al., 2023; Shao et al., 2024; DeepSeek-AI, 2025). These developments have motivated the broader hypothesis that reasoning may also improve alignment and safety by enabling models to deliberate before producing a final response. Several recent works explicitly build on this intuition. Prominent examples include Constitutional AI (Bai et al., 2022) and deliberative alignment methods (Guan et al., 2024), which train models to reason over explicit safety principles before making safety-critical decisions. Following these developments, an important question in the community has been whether safety-trained models can perform robust safety deliberation across diverse settings, and more broadly, how we can make reasoning robust enough to handle diverse and adversarial scenarios.

**Jailbreaks and Robustness of Safety Alignment.** Although our work does not directly study jailbreak attacks on language models, such attacks are closely related to the question of whether reasoning models can robustly deliberate about safety. A large body of work has shown that safety alignment can be bypassed through carefully designed prompting strategies. Early works demonstrated that instruction-tuned models are vulnerable to jailbreak attacks based on role-play prompting, contextual reframing (e.g., hypothetical or past-tense requests), and prompt injection attacks (Wei et al., 2023; Perez & Ribeiro, 2022). Subsequent works developed many-shot in-context jailbreaks, multi-turn conversational attacks, and optimization-based jailbreak strategies (Anil et al., 2024; Hughes et al., 2024; Li et al., 2024; Liu et al., 2024; Zou et al., 2023; Liu et al., 2023; Yu et al., 2023; Mehrotra et al., 2024). An important open direction is to understand how such attacks interact with large reasoning models that can think. In particular, it remains unclear whether longer thinking enables models to better deliberate about harmful requests and resist adversarial manipulations, or whether these attacks simply steer the model toward harmful behavior early in the thinking process itself. Our work takes a first step toward this question by studying how refusal/compliance decisions emerge within the internal reasoning traces of large reasoning models.

**Shallow Safety Alignment.** Closely related to our work, several recent studies suggest that safety alignment in instruction-tuned language models can be surprisingly shallow. Qi et al. (2025) show that refusal behavior may depend heavily on a small number of early tokens and can often be bypassed through simple prefix manipulations. Similarly, Zhao et al. (2025c) demonstrate that weakly aligned models can be leveraged to jailbreak stronger models, while Liu et al. (2026) show that adversarial steering of early generations can substantially alter the model's safety behavior. More recently, Yin et al. (2025) study the representation geometry of reasoning models and show that many poorly aligned reasoning models correctly identify harmful prompts and maintain strong refusal intentions throughout most of the thinking process, but experience a sharp drop in refusal scores near the final tokens before response generation.Our work complements these findings by showing that refusal/compliance behavior is already strongly encoded at the beginning of the thinking process itself, and that subsequent thinking often has limited influence on the final decision.

**Mechanistic Studies of Chain-of-Thought Reasoning.** One closely related line of work seeks to mechanistically understand the role of chain-of-thought reasoning in language models and how reasoning traces connect to the model's final response (Baker et al., 2025; Korbak et al., 2025; Emmons et al., 2025). Closely related to our findings are the recent works of Boppana et al. (2026) and Cox et al. (2026), which show that for many easier reasoning problems, the final answer can often be decoded from hidden representations of the initial tokens, suggesting that later reasoning tokens may partially rationalize an already-determined answer. However, these effects become substantially weaker on harder reasoning tasks. Our work provides complementary evidence in the safety setting, where we show that refusal/compliance behavior is already strongly encoded at the beginning of the thinking trace and that later thinking often has limited influence on the final safety decision.

Tangentially related to our work is a growing literature questioning the faithfulness of chain-of-thought reasoning, showing that generated reasoning traces need not faithfully reflect the internal computation responsible for the model's final answer (Turpin et al., 2023; Lanham et al., 2023; Arcuschin et al., 2025; Boppana et al., 2026). Consistent with these observations, our findings in §2.3 show that even when individual segments of the thinking trace appear deliberative, the model often remains strongly biased toward its initial decision.

# B. Supplementary Details

**Broader Impacts.** This work aims to improve the reliability of safety mechanisms in large reasoning models by clarifying how refusal and compliance decisions emerge during thinking. A potential positive impact is that these findings may help guide the development of models that better distinguish harmful from benign requests, reducing harmful compliance without unnecessarily refusing safe user requests. However, the analysis could also have negative impacts if misused: showing that safety decisions are often encoded early in the reasoning process, and that current defenses may primarily shift models toward refusal rather than improve safety discrimination, could inform attempts to design stronger jailbreaks or evade existing safeguards. We mitigate these risks by focusing on aggregate empirical measurements over existing models and standard benchmarks, rather than releasing new attack datasets, jailbreak procedures, or model checkpoints. We also frame the results around defensive goals, namely the need for training methods that make thinking traces genuinely useful for safety deliberation rather than post-hoc rationalization.

## B.1. Evaluation

**Benchmarks.** We evaluate safety behavior on both harmful and benign prompt sets, so that improvements in harmful-request refusal can be distinguished from increases in over-refusal. For harmful prompts, we use WildJailbreak and FORTRESS. **WildJailbreak** is a large-scale safety dataset containing both direct harmful requests and adversarial jailbreak-style prompts, together with contrastive benign prompts that resemble harmful requests but lack harmful intent (Jiang et al., 2024). We use 2,000 samples from the adversarial-harmful evaluation split, which is designed to test model robustness to jailbreak-style attacks. **FORTRESS** is a frontier-risk safety benchmark focused on national-security and public-safety risks, with expert-crafted adversarial prompts covering CBRNE, political violence and terrorism, and criminal or financial illicit activity (Knight et al., 2025). We use the full dataset to evaluate compliance with high-risk requests.

For benign prompts, we use OR-Bench-Hard, FalseReject, CoCoNot, PHTest, and ORFuzzSet. **OR-Bench-Hard** is the hard subset of OR-Bench, an over-refusal benchmark constructed to evaluate whether models incorrectly reject safe prompts that are likely to trigger conservative refusal behavior (Cui et al., 2025). **FalseReject** contains seemingly toxic but benign queries with structured responses, designed to evaluate contextual safety and unnecessary refusal in settings where surface-level safety cues can be misleading (Zhang et al., 2025c). **CoCoNot** evaluates contextual noncompliance, covering cases where models should or should not refuse depending on the broader context of the request (Brahman et al., 2024). We use its contrast test split as a benign over-refusal set. **PHTest** consists of pseudo-harmful prompts, such as benign requests that contain safety-sensitive words, and is designed to measure false refusals under prompts that superficially resemble harmful requests (An et al., 2024). **ORFuzzSet** is produced by an evolutionary testing framework for over-refusal and contains benign prompts that transfer across models as effective triggers for erroneous safety refusals (Zhang et al., 2025a).

Dataset-provided reference responses, when available, are not used for labeling. For ASR, a harmful prompt is counted as successful if the model complies. For ORR, a benign prompt is counted as over-refused if the model refuses.

**Decoding Settings.** We evaluate each target model using the default recommended sampling configuration for both the base model and all defense-tuned variants of that model. All inference is performed with vLLM (Kwon et al., 2023) in BF16, with a per-prompt budget of 2,048 new tokens for both ASR and ORR splits. These budgets are large enough that fewer than 1% of completions are length-truncated. GPT-OSS-20B is generated in its native harmony channel format; before guardrail scoring, we strip all content outside the final channel. Table 5 lists the exact decoding settings.

**Youden's $J$ Threshold.** AUROC evaluates whether final-refusal examples receive higher probe scores than final-compliance examples, but it does not choose a binary decision threshold. To report balanced accuracy, we must convert the probe's refusal probability into a refusal/compliance prediction. A fixed cutoff such as $0.5$ is arbitrary and can be misleading under class imbalance, since the probe may be well-ranked but poorly calibrated. We therefore choose the cutoff using Youden's $J$, defined for a cutoff $c$ as

$$J(c) = \text{sensitivity}(c) + \text{specificity}(c) - 1,$$

where sensitivity is the refusal-match rate and specificity is the compliance-match rate. Maximizing $J$ is equivalent to maximizing balanced accuracy on the threshold-selection data.

In our setting, the logistic probe is evaluated with 5-fold stratified cross-validation, producing one out-of-fold refusal probability for every example. For each held-out fold $k$, we select the Youden-$J$ cutoff using only the out-of-fold

| Model | Temperature | Top-$p$ | Top-$k$ | Source |
|---|---|---|---|---|
| Qwen3-8B | 0.6 | 0.95 | 20 | Qwen3 release defaults |
| OLMo-3-7B-Think | 0.6 | 0.95 | – | OLMo-3 release defaults |
| Phi-4-Reasoning (14B) | 0.8 | 0.95 | 50 | Phi-4 reasoning defaults |
| GPT-OSS-20B | 1.0 | 1.0 | – | GPT-OSS defaults |

*Table 5.* Decoding settings for target-model generation. All four targets use nucleus sampling with `max_new_tokens` set to 2,048 for ASR and ORR. We use `max_model_len`=16,384 for the 8B/14B models and 8,192 for GPT-OSS-20B. Settings are taken from each model's published generation defaults; we do not tune them per benchmark or per defense. "–" indicates that the parameter is unset, in which case vLLM uses its default.

probabilities from the other four folds, and then apply that cutoff to fold $k$. This ensures that each example is evaluated with both a probe score and a cutoff selected without using that example. We then pool the resulting binary predictions across folds and report refusal-match, compliance-match, and their average as balanced accuracy.

**Thinking Trace Stance Annotation.** We use the prompt template in Figure B.1 to analyze how the visible thinking trace evolves across models. As described in §2.4, we prompt GPT-5.4 to assign stance labels to sentences indicating whether they lean toward compliance, refusal, or neither.

**Measuring Statistical Significance with $p$-Values.** Given a set of variables $x_1, \ldots, x_n$, we begin with the null hypothesis that all variables are drawn from the same underlying distribution and are not statistically distinguishable from one another.

To test this hypothesis, we first standardize the values using their $z$-scores:

$$z_i = \frac{x_i - \mu}{\sigma},$$

where $\mu$ and $\sigma$ denote the empirical mean and standard deviation of the set:

$$\mu = \frac{1}{n}\sum_i x_i, \qquad \sigma = \sqrt{\frac{1}{n}\sum_i (x_i - \mu)^2}.$$

We then compute a two-sided p-value for each variable:

$$p_i = 2\big(1 - \Phi(|z_i|)\big),$$

where $\Phi$ denotes the cumulative distribution function (CDF) of the standard Gaussian distribution.

Finally, we test whether any value is statistically significant at level $\alpha = 0.05$. If there exists any $i$ such that

$$p_i < 0.05,$$

we reject the null hypothesis and conclude that at least one value in the set is statistically different from the others. Otherwise, we fail to reject the null hypothesis, suggesting that the observed variations are consistent with random fluctuations.

### B.2. Inference-Time Defenses

**SafePath ZS (Jeung et al., 2025).** We use the inference-only zero-shot variant, which requires no parameter updates and applies only a prompt-side intervention. After applying the target model's chat template and emitting an open thinking block, `<think>\n`, we prepend a single five-word safety primer, "*Let's think about safety first.*," as the first content of the trace before sampling resumes. Generation is served by vLLM with a maximum of 40,960 tokens.

Prompt Template for Thinking Trace Annotation with GPT-5.4

```
You are analyzing a reasoning trace from an AI model.

  For each chunk assign one label:
  +1  model is planning to comply, engaging with the request, or showing willingness to
     answer
   0  model is neutral, evaluating, restating, or doing general reasoning without clear
     stance
  -1  model is leaning toward refusal, raising safety concerns, or treating the request
     as disallowed

  TRACE:
  Model: {model}
  Prompt: {prompt}
  Chunks: {n_chunks}

  {trace}

  Reply with JSON only--no markdown, no explanation:
  {{"labels": [list of integers], "summary": "one-sentence stance arc"}}
```

**PSR (Phan et al., 2025).** We port the original token-loop PSR procedure to a vLLM backbone while preserving its checkpoint-and-backtrack semantics. Generation is divided into *interval* chunks of $K{=}32$ tokens. After each chunk, we issue a second deterministic completion call with $t{=}0.0$ and max-tokens $= 8$, whose prompt is the trace so far concatenated with the reflection template: "*Let's check if the generated text is harmful or harmless. Is the generated text harmful? Answer:* ". We parse the reflection answer for unsafe, harmful, or dangerous patterns, with explicit handling of negated forms such as *not harmful* and *harmless*. On an unsafe verdict, we discard the $K$-token chunk and resample from the previous checkpoint; on a safe verdict, we commit the chunk and advance the checkpoint. We allow up to $N{=}4$ reflection rounds per prompt, apply reflection only inside the `<think>` span via `--reflect-in think`, disable logit penalties on previously discarded tokens with `logit_bias=0`, and use the same per-target sampling settings and budget as SafePath ZS. Reflection prompts are auxiliary calls only and never enter the user-visible output.

**SafeRemind (Kim et al., 2026).** We implement the entropy-triggered reminder-injection variant on top of vLLM by chunking generation into 64-token windows and requesting the top-20 logprobs for each generated token. From these logprobs, we approximate per-token entropy $H_t$, treating the residual probability mass outside the top-20 as zero following the original implementation. Inside the `<think>` block, we trigger on a local entropy drop using a sliding window of $W{=}10$ tokens with threshold $H_t < \mu_W - \alpha\sigma_W$ and $\alpha{=}1.5$. On a trigger, we commit the trace up to and including the trigger token, append the fixed reminder phrase, "*Wait, let me think again. Does this request involve potentially harmful content?*," and continue generation with a cooldown of $W$ tokens before the next trigger is eligible. We cap the procedure at 3 injections per prompt and use the *fixed*-phrase mode; the optional auto-reminder mode, which uses an LLM to generate a query-specific reminder, is not used in our runs. Sampling settings and generation budget match the other two defenses.

### B.3. Training-Based Defenses

We train each defense on four target models spanning three reasoning-model families and two parameter scales: Qwen3-8B, OLMo-3-7B-Think, Phi-4-Reasoning (14B), and GPT-OSS-20B. All runs use bfloat16 on NVIDIA H100-80GB GPUs. The 8B/14B models are trained on a single GPU, while GPT-OSS-20B uses three-way DeepSpeed ZeRO-3 with gradient checkpointing. For each defense, we keep the upstream paper's data, recipe, and optimizer settings fixed across target models, varying only the family-specific chat template and resource allocation.

**STAR-1.** STAR-1 (Wang et al., 2025b) performs full-parameter SFT on a 1,000-example corpus of safety-reasoning traces, consisting of PKU-SafeRLHF(Ji et al., 2024) prompts paired with GPT-4-style reasoning rationales. We follow the released configuration: 5 epochs, AdamW with $\beta_1 = 0.9$ and $\beta_2 = 0.95$, weight decay $10^{-4}$, learning rate $1 \times 10^{-5}$, cosine schedule with 5% linear warmup, effective batch size 128 (per-GPU batch size 1 with 128 gradient-accumulation steps),

maximum sequence length 8,192, seed 2002, and gradient checkpointing.

**SafeKey.** SafeKey (Zhou et al., 2025b) augments STAR-1 with sentence-level annotations identifying the "key sentence" that pivots the thinking trace toward a safe completion. The annotations include `sentence_index`, `summary_end_idx`, and `next_sentence_end_idx`; the released mixture contains 1,915 examples. Training adds two auxiliary objectives on top of the SFT loss: a binary safety-head classifier and a key-sentence-prediction head, enabled via `-safety_head` and `-key_sentence_prediction`. Optimizer settings match STAR-1: learning rate $1 \times 10^{-5}$, weight decay $10^{-4}$, $\beta = (0.9, 0.95)$, cosine schedule with 5% warmup, effective batch size 128, and maximum sequence length 8,192. We train for 10 epochs, with the auxiliary heads active during the final 4 epochs as in the released code.

**R1-ACT.** R1-ACT (In et al., 2025) is a LoRA SFT recipe trained on a 959-example corpus of harmful instructions paired with early-refusal traces of the form "Okay, ... this instruction is harmful ... no further consideration is necessary." We use the released hyperparameters: LoRA rank 16, $\alpha = 16$, learning rate $1 \times 10^{-5}$, 5 linear warmup steps, effective batch size 16 (per-GPU batch size 1 with 16 gradient-accumulation steps), 15 epochs, maximum sequence length 8,192, ZeRO-2, and adapter merging after training.

**STAIR.** STAIR (Zhang et al., 2025b) uses a two-stage curriculum. Stage 1 is a LoRA SFT warm start on 20,000 examples, consisting of 10K PKU-SafeRLHF examples and 10K UltraFeedback (Cui et al., 2023) examples. We use LoRA rank 64, $\alpha = 128$, learning rate $2 \times 10^{-5}$, 10 warmup steps, 3 epochs, effective batch size 16, maximum sequence length 4,096, and ZeRO-2. Stage 2 performs three rounds of MCTS-driven safety preference learning. In each round, the current policy samples MCTS rollouts, extracts chosen/rejected pairs, and runs LoRA DPO on the merged adapter with $\beta = 0.1$, learning rate $5 \times 10^{-6}$, 10 warmup steps, 1 epoch, LoRA rank 64, $\alpha = 128$, maximum sequence length 4,096, maximum prompt length 1,024, and effective batch size 16. The final adapter is merged after the third round.

**ThinkSafe.** ThinkSafe (Lee et al., 2026) applies a forward-KL distillation objective during LoRA SFT. For each target model, we synthesize data by rolling out the base model on SafeChain (Jiang et al., 2025) prompts using vLLM with $T = 0.6$, $p = 0.95$, and top-$k = 20$. WildGuard filters unsafe completions, keeping refusals for harmful prompts and helpful answers for benign ones; we also inject a refusal-style prefix into the harmful subset. Training incorporates a forward-KL term between the LoRA-adapted policy and frozen base model. We train for 1 epoch with learning rate $1 \times 10^{-5}$, cosine decay with 10% warmup, effective batch size 8, maximum sequence length 4,096, LoRA rank 32, $\alpha = 16$, and gradient checkpointing. Family-specific targets follow the upstream code: `q_proj/k_proj/v_proj/o_proj` for GPT-OSS, `qkv_proj/o_proj/gate_up_proj/down_proj` for Phi-4, and all-linear targets for Qwen3/OLMo.

**RAPO.** RAPO (Wei et al., 2026) is a two-stage SFT→GRPO pipeline that first elicits a controlled safety-reasoning prefix and then optimizes a risk-aware reward. For the SFT stage, we generate 800 prompts using the upstream mixture of the STAR-1 benign split (400) and Strata-Sword (400) (Zhao et al., 2025b), and apply the upstream two-pass prompting procedure. Pass 1 produces a safety-reasoning trace with `ADAPTIVE_THINKING_SYSTEM_PROMPT`; pass 2 prefills that trace and generates a controlled response. We then run full-parameter ZeRO-3 SFT on the resulting prompt-response pairs for 3 epochs using the upstream defaults, including per-device batch size 2, no packing, and BF16/FP16 according to the released configuration. For the RL stage, we run GRPO (Shao et al., 2024) with RAPO's risk-aware and general rewards. The reward model is a Qwen3-8B judge prompted with the `REWARD_JUDGE`, `SAFETY_JUDGE`, and `BENIGN_JUDGE` system prompts from the original implementation. We train for 3 epochs with 4 generations per prompt, maximum completion length 2,048, sampling temperature 1.2, top-$p = 0.9$, and an upstream recipe consisting of 300 samples from the WildJailbreak train split, 100 from STAR-1, and 400 from the STAR-1 benign split.

We make the following adaptations for Phi-4-Reasoning and GPT-OSS-20B. RAPO's pass-2 prefix injection assumes the standard ChatML `<think>...</think>` structure of the original Qwen3 backbone, so applying the method to our non-Qwen targets required two minimal infrastructure adaptations, neither of which changes the underlying methodology. For Phi-4-Reasoning, which we find to sometimes fail to emit `</think>`, we force-close `</think>` directly inside the pass-2 prefix at the max thinking budget of 8192 tokens, `<think>\nOkay, {SR}</think>\n\n`, anchoring the model outside the thinking channel so that it produces the final response; we also extend the slicer to consume Phi-4-Reasoning's `<|im_sep|>` separator and apply a loop-truncation regex to the post-`</think>` span as a safety net. For GPT-OSS-20B, which uses the harmony format rather than ChatML, we treat the analysis channel as the safety-reasoning trace and inject the pass-2 prefix so that the trace is placed in the analysis channel before control is handed to the final channel. This anchors the

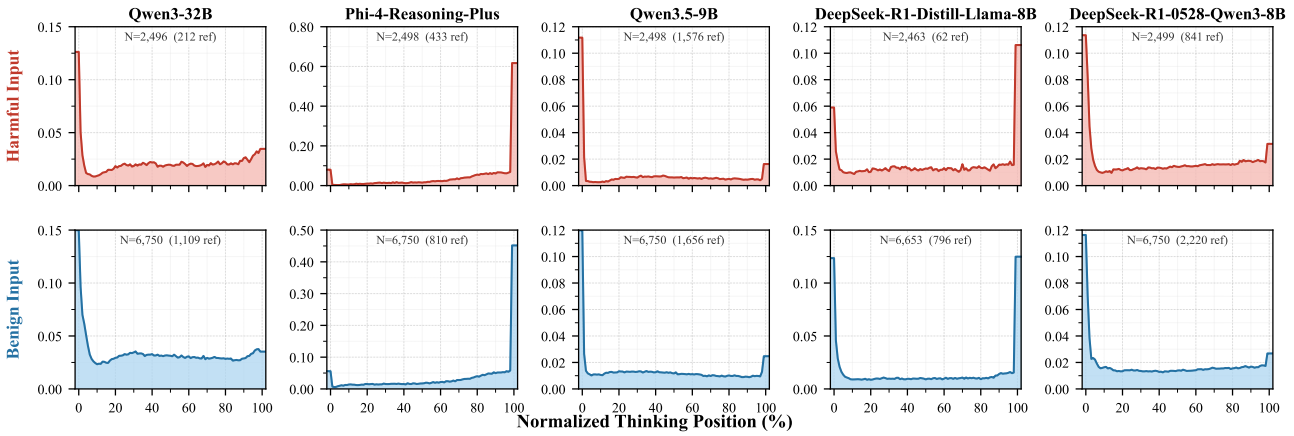


*Figure 5.* Fisher discriminant plot for supplementary thinking models. All five models reproduce the same qualitative valley signature observed in the four primary models, with high pre-decoding separability, a sustained drop through the middle of the trace, and re-crystallization at the end.

model at the start of the final response, so its raw generation already corresponds to the final answer; accordingly, the slicer consumes the raw generated text directly rather than slicing on an assistant boundary that would be disrupted by the prefilled channel transition. Both adaptations preserve RAPO's design intent: using prior-pass safety reasoning as a thinking-trace prefix to elicit a controlled response.

## C. Additional Results

### C.1. Supplementary Fisher Discriminants

We present Fisher discriminant plots for additional thinking models in Figure 5. Qwen3-32B is the 32B-parameter variant of Qwen3 and uses the same thinking-mode chat template as our primary Qwen3-8B model. Phi-4-Reasoning-Plus is an RL-trained successor to Phi-4-Reasoning, post-trained with reinforcement learning from verifiable rewards (RLVR) on top of the SFT-only Phi-4-Reasoning checkpoint. The two DeepSeek-R1 (DeepSeek-AI, 2025) distilled models, DeepSeek-R1-Distill-Llama-8B and DeepSeek-R1-0528-Qwen3-8B, are open-weight models that acquire R1-style thinking through supervised fine-tuning on traces sampled from DeepSeek-R1 rather than through their own reasoning-RL stage.

We observe that all five supplementary models reproduce the same valley signature observed in the primary models. Qwen3-32B shows that the effect persists under a $4\times$ parameter scale-up within the Qwen3 family. Phi-4-Reasoning-Plus exhibits the same first-token and last-token concentration as Phi-4-Reasoning, with a similarly sharp drop in mid-trace separability, suggesting that additional RLVR post-training may not effectively move the safety decision into the visible thinking trace. Both DeepSeek-R1 distilled models also exhibit the valley, indicating that this phenomenon is present in distillation-based models.