Do Thinking Tokens Help with Safety?

Narutatsu Ri [1] **Abhishek Panigrahi** [1] **Sanjeev Arora** [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 spanning GPT-OSS, Qwen, Olmo, and Phi families, we find that the eventual refusal/compliance outcome is already strongly readable before any visible thinking, with a probe on the first token's hidden representation predicting refusal/compliance with 0.84–0.95 AUROC and $\sim 88\%$ balanced accuracy. Here, thinking turns out to behave more like prefix completion than deliberative revision, with the final outcome rarely changing after the first $\sim 20\%$ of thinking. Inspecting these thinking traces reveal that among segments that appear deliberative at the text level, only a minority affect the final outcome. More strikingly, $\sim 74\%$ of text-level deliberations occur when the response distribution is already locked to one refusal/compliance side, even as the trace continues to look deliberative. We also find that existing inference-time and training-based safety interventions, despite being motivated by the goal of activating deliberation, largely shift model behavior toward over-refusal while suppressing already scarce deliberation signals. Our results suggest that safety behavior in current reasoning models is much less deliberative than commonly assumed, and highlight the need for methods that induce real safety deliberation.[1]

[1]Princeton University, Princeton, NJ, US. Correspondence to: Narutatsu Ri <nr3764@princeton.edu>.

Trustworthy AI for Good (AI4GOOD) Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

[1]Code can be found at `github.com/narutatsuri/lrm_safety_deliberation`.

1. Introduction

Large reasoning models (LRMs) are trained to "think" before answering by generating a long reasoning trace before producing a final response. This use of inference-time compute has led to substantial capability gains on many verifiable tasks, such as math and coding (Wei et al., 2022; Kojima et al., 2022; Zhou et al., 2023; Shao et al., 2024; DeepSeek-AI, 2025). Excitingly, the broader appeal of reasoning is that these capability gains appear to extend beyond verifiable tasks, with growing evidence of improvements in 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 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. Importantly, a safe and helpful model should simultaneously attain low ASR and ORR (Kim et al., 2025). A natural belief is that thinking should improve this trade-off, as models may use the reasoning process to recognize harmful requests while preserving helpfulness on benign inputs. This intuition is appealing, and indeed 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. Even so, it would be natural to expect reasoning models to behave differently for the better; after all, extended thinking often changes and improves answers on many tasks, so the thinking trace should seem likely to contribute directly to better safety decisions. We challenge this assumption by investigating the following question:

> *Does thinking truly improve safety decision making in reasoning models?*

As we detail later, our findings confirm this concern: we find

evidence that safety decisions in reasoning models may be consolidated early on in the reasoning process, and thinking plays a weak role in meaningfully improving safety behavior. Still, one might hope that defenses tailored to reasoning models address this failure by making the trace more safety-aware, as many methods explicitly aim to do (In et al., 2025; Kim et al., 2026). This raises a second question:

> *Do existing defenses meaningfully improve both the ASR–ORR tradeoff and safety deliberation capability of reasoning models?*

Our Contributions. In this work, we provide evidence that safety behavior in current LRMs may be far less deliberative than the reasoning paradigm suggests. First, we present a surprising finding that a model's eventual refusal/compliance outcome is already strongly readable even before any visible thinking content 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 0.84–0.95 and up to 88% balanced accuracy. 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, thinking does not meaningfully contribute to improved safety decisions; instead, once a short thinking prefix is fixed, later continuations tend to preserve the refusal/compliance trajectory established early in the thinking trace (§3.1). Further, while thinking traces do exhibit sentence-level patterns that resemble safety deliberation at the text level, these segments turn out to have minimal influence on the final response (§3.2).

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, while also suppressing the already weak deliberation signals (§4.1). This pattern is consistent with the prefix-continuation behavior observed in §3.1, where later thinking rarely steers the model away from early refusal/compliance trajectories. In contrast, training-based methods induce larger behavioral shifts, but do so by pushing models toward the same over-refusal pattern and do not meaningfully increase deliberation in the thinking trace (§4.2). Our findings highlight the need for new training objectives that reward true deliberation rather than merely encouraging models to generate safety-flavored reasoning.

2. Safety Behavior of Reasoning Models

2.1. Setup

Notation. Given a prompt $\mathbf{x} = (x_1, \ldots, x_P)$, a language model autoregressively generates a completion $\mathbf{y} = (y_1, \ldots, y_T)$. For the reasoning models we study, model-specific control tokens (e.g., `<think>...</think>`, `<|channel|>analysis<|message|>...<|end|>`) 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 τ denotes the number of tokens used in the thinking block. 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 $t \in \{0, \ldots, 99\}$. We will frequently use $\mathbf{h}_0$ to denote the final-layer hidden representation of the first token r_1 in the thinking trace.

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 measuring ASR and 2,885 benign prompts for ORR, with dataset composition shown in Table 4. For analyses requiring finer-grained estimates on benign prompts, we increase the benign pool to 6,750 instances by additionally using PHTest (An et al., 2024) and ORFuzzSet (Zhang et al., 2025a). Refusal/compliance labels are assigned only to $\mathbf{a}$, not to the intermediate thinking trace $\mathbf{r}$. 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 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.[2]

2.2. Decisions are Readable Before Visible Reasoning

We first show that the representation of the first token in the thinking trace encodes a refusal/compliance signal that corresponds surprisingly well to the final decision. We do so by probing the directional geometry of last-layer hidden states at each token on refusal and compliance thinking traces separately. For each model, we generate one completion per prompt in the full ASR pool and expanded ORR pool of §2.1, label each completion's final response using the four-guardrail vote, and separate instances into refusal and

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

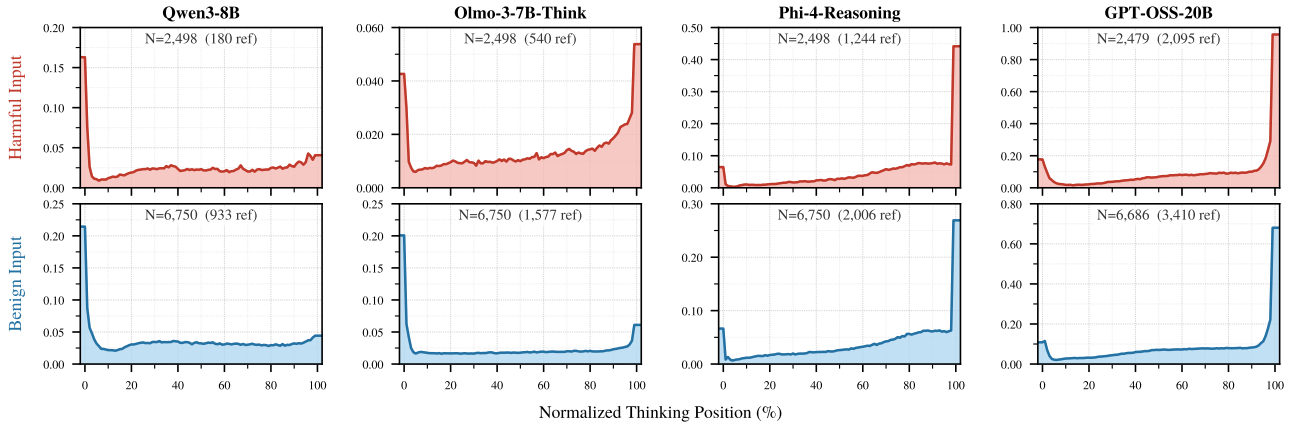

Figure 1. Fisher discriminant ratio $J(t)$ between eventual refusal and compliance traces, plotted against normalized position t in the thinking trace. Across all models, $J(t)$ traces the same U-shape (the refusal valley): high at the first generation position, dropping through the body of the thinking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

compliance groups within each pool.

We consider two measures at each position t between refusal and compliance groups. First, we use a trace-normalized Fisher discriminant ratio to compute the distance between the two groups' mean representations, which measures group distinguishability modulo within-class variability:

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

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 $\operatorname{tr} \boldsymbol{\Sigma}_W(t)$ is the trace of the within-class covariance $\boldsymbol{\Sigma}_W(t)$, i.e., the within-class variance. Second, we train an ℓ_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 standardize each feature dimension and project onto the top 100 principal components of $\mathbf{h}_0$, fitting both within each training fold of a 5-fold stratified cross-validation that yields one out-of-fold refusal probability for every example. We report two metrics on these held-out probabilities: AUROC, a threshold-free rank-order measure of class separability, and balanced accuracy (BAcc), a thresholded measure of probe accuracy. See §B for additional details.

Refusal/Compliance is Strongly Readable at the First Token Representation. See Figure 1. Observe that $J(t)$ shows a notable *valley* shape that begins high at $t{=}0$, drops through the body of the thinking 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. This behavior is also robust across various reasoning models; we include plots for additional models in §C.2.

Most strikingly, this early readout closely matches the refusal/compliance outcome of the model's final response. As shown in Table 1, the probe on the first token representation achieves high AUROC across all models (0.840–0.948), indicating that the two outcomes are linearly separable at $t{=}0$. Complementing this, the probe also achieves a balanced accuracy (BAcc) of 0.763–0.878, 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.

Readout Signal Spikes at the End of Prefill. Because the first thinking-token representation is computed after the prompt has already been processed, this raises the question of where the signal emerges across the input token representations during prefill. We next examine this by applying the probing approach from §2.2 to the prefill. We first normalize positions from the user query through the end of the input[3], and train individual probes for 20 equally spaced positions.

See Figure 2 for results. We find that the probing readout is weak throughout most of the prefill but rises sharply near the end across all models, suggesting a late-prefill consolidation of the refusal/compliance outcome that visible thinking subsequently inherits. This transition is most pronounced under the Fisher discriminant, with AUROC and BAcc exhibiting a slight upward trend before also spiking at the end.

3. Utility of Thinking

Importantly, the early readout result in §2.2 should not by itself be considered a failure mode of reasoning models, as certain prompts may be unambiguous enough that the

[3]We remove chat-template scaffolding as it is the same across inputs.

Table 1. First-token probe AUROC and balanced accuracy (BAcc) values by probe type. Rep. uses $\mathbf{h}_0$, and Text uses the first token's TF-IDF feature. A linear probe on $\mathbf{h}_0$ predicts the final refusal/compliance outcome, while text-only controls have less predictive power.

Model	Probe	Harmful Pool		Benign Pool		Combined	
		AUROC	BAcc	AUROC	BAcc	AUROC	BAcc
Qwen3-8B	Rep.	0.948 ± 0.010	0.878 ± 0.019	0.888 ± 0.025	0.804 ± 0.031	0.971 ± 0.004	0.919 ± 0.008
	Text	0.492 ± 0.042	0.502 ± 0.005	0.502 ± 0.019	0.502 ± 0.002	0.499 ± 0.017	0.502 ± 0.002
Olmo3-7B-Think	Rep.	0.881 ± 0.016	0.794 ± 0.017	0.891 ± 0.012	0.810 ± 0.014	0.949 ± 0.005	0.872 ± 0.008
	Text	0.500 ± 0.000	0.500 ± 0.000	0.500 ± 0.000	0.500 ± 0.000	0.500 ± 0.000	0.500 ± 0.000
Phi-4-Reasoning	Rep.	0.875 ± 0.014	0.792 ± 0.016	0.840 ± 0.016	0.763 ± 0.018	0.897 ± 0.006	0.816 ± 0.008
	Text	0.561 ± 0.022	0.537 ± 0.011	0.502 ± 0.014	0.516 ± 0.004	0.515 ± 0.013	0.518 ± 0.004
GPT-OSS-20B	Rep.	0.916 ± 0.013	0.829 ± 0.020	0.878 ± 0.012	0.796 ± 0.013	0.939 ± 0.004	0.861 ± 0.007
	Text	0.678 ± 0.031	0.671 ± 0.025	0.619 ± 0.013	0.597 ± 0.009	0.601 ± 0.012	0.592 ± 0.008

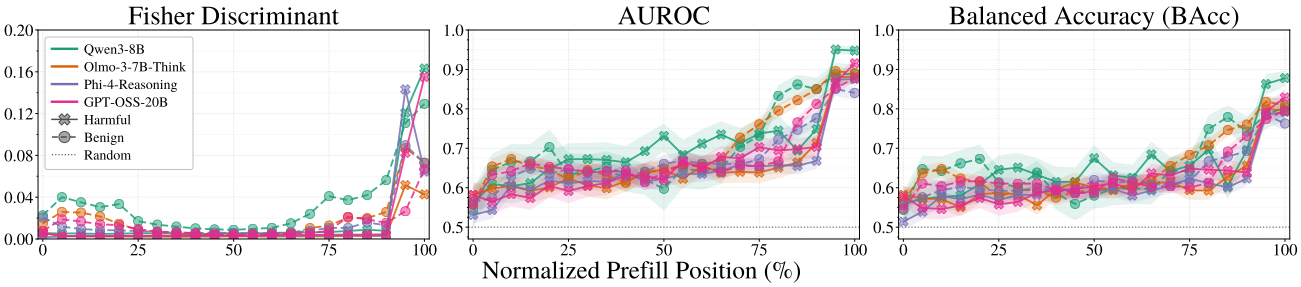

Figure 2. Prefill readout dynamics for the refusal/compliance outcome. We report the Fisher discriminant, AUROC, and BAcc for linear probes trained at 20 normalized positions from the user query to the end of prefill. For all three metrics, the readout remains weak over most of the input and spikes near the final prefill positions across all models and both pools.

correct decision can be made trivially. In general, however, the thinking process should be able to affect the model's final behavior, rather than simply rationalize a decision already evident from the early representations. Here, we investigate the extent to which this is possible in current models.

3.1. How Does Thinking Contribute to the Decision?

A follow-up question from §2.2 is what role the thinking trace plays in the decision. Here, we provide evidence that for current models, it largely serves as post-hoc rationalization. For each prompt, we sample $M{=}8$ independent thinking traces and form prefixes at budgets $B \in \{20\%, 40\%, 60\%, 80\%, 100\%\}$ of the natural trace length by truncating each trace at the corresponding fraction. We sample $N{=}8$ final responses conditioned on each prefix and label the responses accordingly with guardrails. Given a fixed prompt and budget B, let $z_{m,B,n} \in \{0, 1\}$ denote whether the n-th final response from sampled thinking prefix m with budget B is labeled as a refusal, and let $q_{m,B} = \frac{1}{N}\sum_n z_{m,B,n}$ be the refusal rate after conditioning on that prefix. We measure the following form of variance, which captures how frequently different continuations from the same prefix lead to different outcomes:

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

where the factor of 4 normalizes Bernoulli variance so that the maximum possible variance is 1. We pool results across both ASR and ORR pools and examine per-model results for each budget B.

Thinking Serves as Prefix Completion Rather than Prefix Adjustment. See Figure 3. Observe that continuation variance is already near zero even at $B{=}20\%$ (<0.2), and decreases further as the prefix length grows across all models. This indicates that independent continuations generally preserve the decision trajectory set by the prefix even for very short prefixes, and later thinking rarely steers away from this decision.[4]

These results indicate that thinking tokens may not serve as the deliberative space we might expect, raising the question of whether thinking is useful to begin with. We study this by examining whether thinking improves the final decision relative to the no-think ($B = 0\%$) setting. For each input in the ASR and ORR pools, we sample $K{=}32$ independent generations under each setting. We label final responses via a four-guardrail majority vote, and assign the input's response as refusal if more than $K/2$ rollouts are refusals and compliance otherwise. A *majority-label flip* is then the input-level label change between the no-think and think settings, and desirable flips are ones that change compliance to refusal on the ASR pool and vice versa for ORR.

[4]We also examine whether forcing additional test-time compute beyond the natural stopping point changes this behavior, and find that it does not; see §C.3 for details.

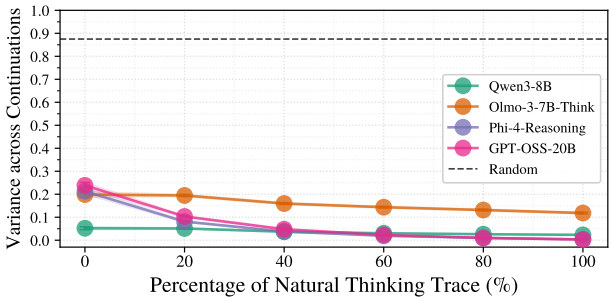

Figure 3. Continuation variance with fixed prefix. Dashed line indicates fair-coin reference (0.875, n=8; expected variance given coin flip). Observed values lie far below baseline.

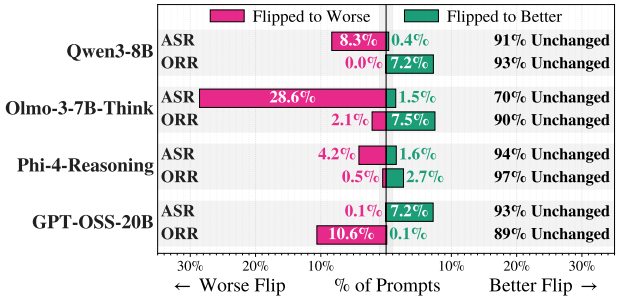

Figure 4. Majority-label flips induced by enabling thinking vs. no thinking. Green bars indicate flips toward desired behavior. Red bars indicate flips in undesired direction. At most 30% of prompts flip under thinking, and across all models these flips improve either ASR or ORR but not both.

Turning Thinking On Does Not Improve the ASR–ORR Tradeoff. See Figure 4. First, observe that thinking modifies a small fraction of the inputs across model-pool pairs, and 70–97% of prompts retain the same label. More crucially, no model shows a consistent improvement on both sides of the ASR–ORR tradeoff with thinking: Qwen3-8B, Olmo-3-7B-Think, and Phi-4-Reasoning tend to over-comply while GPT-OSS-20B exhibits the opposite tradeoff, which indicates that thinking largely shifts models along the ASR–ORR tradeoff rather than improving both.[5]

3.2. Do Individual Sentences Steer the Decision Meaningfully?

§3.1 suggests that the thinking trace often has limited effect on an outcome already readable early on in the generation. Then, what function does the text in the thinking trace serve?

To quantify this, we analyze safety deliberation at the sentence level. For 250 randomly sampled prompts each from the ASR and ORR pool (500 prompts in total), we sample M=4 independent thinking rollouts and segment each into sentences $c_{1:T}$. We use three frontier models (GPT-5.4, Gemini 3 Pro, and Sonnet 4.6) to assign each sentence a stance label in $\{-1, 0, +1\}$, corresponding to refusal-leaning, neutral, and compliance-leaning text, respectively.[6]

We first measure the fraction of rollouts that exhibit at least one *oscillation*, i.e., a change in stance from refusal-leaning to compliance-leaning or vice versa, taken over all M rollouts of all prompts. On rollouts where we find oscillations, we then measure whether each oscillation has a statistically significant effect on the final response. For each oscillation, let i_{pre} denote the index of the last sentence with the previous stance, and let i_{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 $x^{\text{post}} = c_{1:i_{\text{post}}}$. We estimate the refusal rates at $q_{x^{\text{pre}}}$ and $q_{x^{\text{post}}}$ by truncating the thinking trace at each and using K=100 final-response generations. The effect of the segment between i_{pre} and i_{post} is then quantified using its *stance-aligned shift*:

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

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.

Depending on the values of $q_{x^{\text{pre}}}$ and $q_{x^{\text{post}}}$, we characterize each oscillation along two axes. First, the oscillation is "locked" if $q_{x^{\text{pre}}}$ is saturated (≤ 0.05 or ≥ 0.95), i.e., the majority of K rollouts already collapse to one label. Second, for each oscillation we measure whether it imposes a statistically significant change in the final response distribution between x^{pre} and x^{post}. Statistically significant oscillations are termed *meaningful*, while others are denoted as *performative*. See §B for additional evaluation details.

Most Thinking Trace Segments Show No Significant Effect. See Figure 5. First, observe that only a minority of rollouts contain any oscillation (15–34%). Among those that do, the large majority of swings (71–92%) are performative, with no statistically significant shift in the final refusal/compliance outcome. In fact, 72.8–76.6% of all oscillations occur while the outcome is already locked, meaning that the stance label changes at the text level while the model's final decision remains unchanged. Nevertheless, we find that 87–98% of segments that do exert a significant effect usually do so in the direction implied by their stance, hinting that the desired behavior is rare but present.

4. What Do Existing Defenses Do?

The results in §2 suggest that safety outcomes are often strongly readable from the first token's representation,

[5] We also consider more permissive measures, e.g., labeling thinking as beneficial whenever the rollout-level refusal rate shifts in the desired direction. It turns out the same conclusion holds even under such lenient criteria. See §C.3 for further details.

[6] See §C.3 for an evaluation of inter-annotator agreement among frontier models.

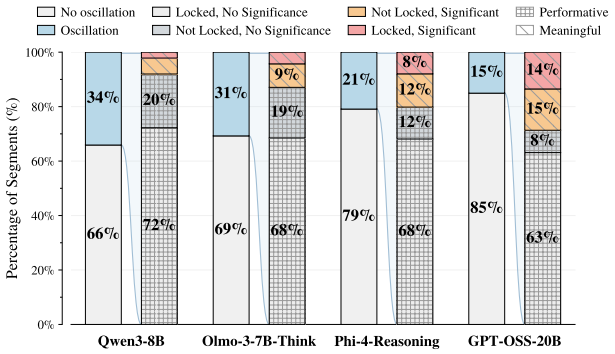

Figure 5. Effect of individual sentence segments on the final refusal/compliance decision. Main bar: fraction of $M{=}4$ thinking rollouts that contain at least one oscillation. Expanded bar: every oscillation segment classified along locked ($q_{x^{\text{pre}}}$ satisfying ≤ 0.05 or ≥ 0.95) vs. not locked, and meaningful vs. performative. Across all four models, the large majority of segments are performative. Results are pooled across ASR and ORR pools; see §C.4 for a per-pool breakdown.

and that additional thinking does not reliably improve both ASR and ORR. We now ask whether existing safety defenses change this picture along two axes: whether they improve the ASR–ORR tradeoff, and whether they increase the influence of thinking traces as measured by sentence-level oscillation analysis. We examine two classes of practical defenses: inference-time interventions and training-based defenses based on supervised fine-tuning (SFT) or reinforcement learning (RL).

4.1. Inference-Based Defenses

Inference-time defenses provide a direct test to whether LRM safety can be improved by simply intervening on the thinking trace itself. Most existing defenses are of this form, and do so by inserting (off-policy) 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; **SafeRemind** (Kim et al., 2026) dynamically inserts safety-reminding phrases during thinking when entropy indicates a decision-locking point.[7] The common premise is that injecting safety considerations into the trace should "activate" the model's safety deliberation capabilities and steer subsequent reasoning (Kim et al., 2026).

However, §3 suggests a more pessimistic view: If LRMs do not effectively use thinking for safety deliberation, such interventions should either have little effect (§3.2) or, at best, push models toward over-refusal by forcing them to complete a thinking block that cautions against compliance (§3.1). We thus evaluate ASR and ORR for all defenses and report them in Table 2. The results are consistent with this hypothesis: inference-time defenses either lower ASR at the cost of higher ORR or have limited effect. 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.

We further examine what happens at the sentence-level compared to the base model; do non-performative segments increase under inference-time defenses? We apply the analysis from §3.2 ($M{=}4$ traces, 250 instances from ASR/ORR pools) to inference-time defenses and find that they do not; see Table 3. Notably, both the average number of total oscillations per trace and meaningful oscillations decrease sharply, with drops up to 95% for Phi-4-Reasoning and over 60% on both metrics for Olmo-3-7B-Think under all three defenses. For these two models, meaningful oscillation counts decrease at least as fast as total oscillation counts. Overall, this suggests that inference-time defenses may be a less promising lever for safety deliberation in reasoning models, as they improve ASR by suppressing already sparse deliberation rather than making the trace more deliberative.

4.2. 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 (RL)-based 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.

[7] See Appendix B.2 for further details on implementation.

Table 2. ASR and ORR of inference-time and training-time defenses under four-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.

		Qwen3-8B		Olmo-3-7B-Think		Phi-4-Reasoning		GPT-OSS-20B	
	Method	ASR ↓	ORR ↓	ASR ↓	ORR ↓	ASR ↓	ORR ↓	ASR ↓	ORR ↓
	Base	86.9	6.0	70.6	21.2	47.3	38.5	26.5	44.9
Inf.	PSR	86.90 ↑0.0	6.10 ↑0.1	38.00 ↓32.6	43.50 ↑22.3	32.70 ↓14.6	70.60 ↑32.1	29.80 ↑3.3	42.70 ↓2.2
	SafeRemind	80.80 ↓6.1	11.00 ↑5.0	32.50 ↓38.1	44.90 ↑23.7	33.30 ↓14.0	70.40 ↑31.9	21.30 ↓5.2	49.20 ↑4.3
	SafePath-ZS	68.20 ↓18.7	13.50 ↑7.5	30.70 ↓39.9	49.10 ↑27.9	31.90 ↓15.4	71.90 ↑33.4	22.50 ↓4.0	48.80 ↑3.9
Training	STAR1	41.0 ↓45.9	45.5 ↑39.5	39.2 ↓31.4	49.8 ↑28.6	35.8 ↓11.5	45.3 ↑6.8	71.7 ↑45.2	14.8 ↓30.1
	SafeKey	56.3 ↓30.6	28.0 ↑22.0	48.1 ↓22.5	36.9 ↑15.7	52.7 ↑5.4	33.5 ↓5.0	71.0 ↑44.5	19.4 ↓25.5
	R1-ACT	57.0 ↓29.9	37.8 ↑31.8	85.4 ↑14.8	27.7 ↑6.5	53.5 ↑6.2	39.2 ↑0.7	54.6 ↑28.1	46.7 ↑1.8
	ThinkSafe	30.9 ↓56.0	44.6 ↑38.6	46.7 ↓23.9	38.4 ↑17.2	40.8 ↓6.5	40.8 ↑2.3	18.8 ↓7.7	52.3 ↑7.4
	RAPO	34.7 ↓52.2	35.0 ↑29.0	17.7 ↓52.9	49.9 ↑28.7	55.4 ↑8.1	53.4 ↑14.9	44.0 ↑17.5	67.5 ↑22.6
	STAIR	53.8 ↓33.1	45.3 ↑39.3	54.6 ↓16.0	38.3 ↑17.1	47.5 ↑0.2	38.8 ↑0.3	26.4 ↓0.1	68.6 ↑23.7

Table 3. Per-trace oscillation counts for the undefended Base model and three inference-time defenses. Osc. is the average number of stance oscillations per reasoning trace. Meaningful denotes the average number of oscillations that significantly shift the final-answer comply rate. Subscripts report percent change relative to Base. Red cells indicate decreases and green cells indicate increases. Across 22 of 24 inference cells, defenses reduce oscillations.

		Qwen3-8B		Olmo-3-7B-Think		Phi-4-Reasoning		GPT-OSS-20B	
	Method	Osc.	Meaningful	Osc.	Meaningful	Osc.	Meaningful	Osc.	Meaningful
	Base	1.20	0.10	1.43	0.19	0.66	0.14	0.32	0.09
Inf.	PSR	0.64 ↓47%	0.07 ↓27%	0.48 ↓66%	0.02 ↓87%	0.03 ↓95%	0.01 ↓93%	0.25 ↓23%	0.09 ↓5%
	SafeRemind	1.13 ↓5%	0.10 ↑6%	0.30 ↓79%	0.03 ↓85%	0.14 ↓78%	0.02 ↓87%	0.25 ↓22%	0.13 ↑38%
	SafePath-ZS	0.72 ↓40%	0.07 ↓29%	0.56 ↓61%	0.03 ↓85%	0.12 ↓82%	0.05 ↓64%	0.17 ↓47%	0.05 ↓41%

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

Training-Based Defenses Do Not Improve the ASR–ORR Tradeoff. From Table 2, we see that as with inference-time defenses, training-based defenses generally shift models along the ASR–ORR tradeoff rather than improving both, with the direction and magnitude depending on the model and method. ThinkSafe and RAPO reduce harmful compliance at the cost of over-refusal, as seen on Qwen3-8B with ThinkSafe (ASR 86.9 → 30.9, ORR 6.0 → 44.6) and Olmo-3-7B-Think with RAPO (ASR 70.6 → 17.7, ORR 21.2 → 49.9). Other methods move in the opposite direction or worsen both axes: STAR1 on GPT-OSS-20B lowers ORR from 44.9 to 14.8 while raising ASR from 26.5 to 71.7, while R1-ACT raises ASR on three of four models and ORR on all four.

Training-Based Defenses Do Not Improve Deliberativeness. Beyond the ASR–ORR tradeoff, we also examine how training reshapes the model's safety deliberation mechanism. We do so by looking at how the first-token probe readability of §2.2 changes for trained models along with the sentence-level oscillation analysis of §3.2.

Figure 6 shows these results and separates methods by training trace source: on-policy methods train with the model's own traces, while off-policy methods distill externally obtained traces. On-policy methods (ThinkSafe, STAIR) turn out to preserve first-token readability but reduce meaningful and performative oscillations alike, further weakening deliberation. Off-policy methods (STAR-1, SafeKey) instead suppress the early refusal/compliance signal (AUROC −0.11) and reduce overall oscillations, but largely preserve meaningful oscillation counts. Overall, training-based defenses can substantially alter model behavior, but neither family consistently improves the ASR–ORR tradeoff or make the visible thinking trace more deliberative.

5. 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 encodes a refusal/compliance decision and that this prediction closely matches the model's final response. We also find that thinking does not consistently improve safety decisions, and seemingly deliberative thinking segments exert limited influence on the

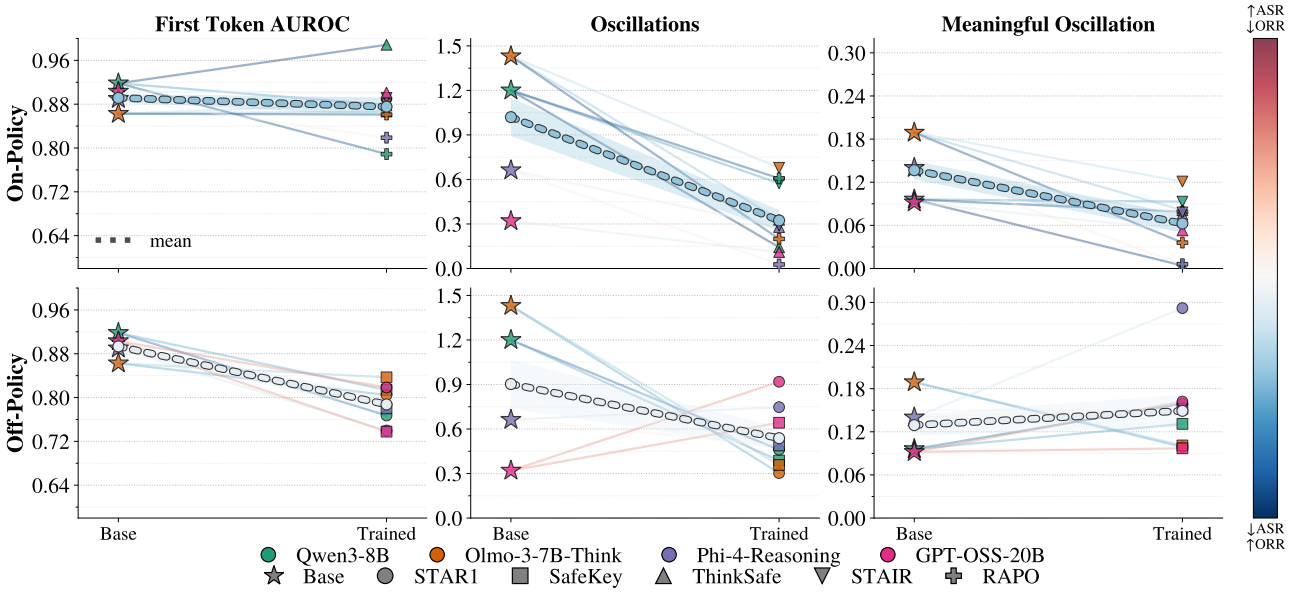

Figure 6. On- and off-policy training reshape reasoning in opposite ways, and neither consistently increases meaningful deliberation. Each line tracks one model-defense pair from Base to its trained variant across first-token readability, stance oscillations per trace and meaningful oscillations per trace. Rows split methods by training trace source. On-policy methods fine-tune on the model's own traces, while off-policy methods distill external traces. Line color indicates whether ASR decreases or increases after training. Dashed lines show per-row means with 95% CI. We exclude R1-ACT, which fails as a defense in Table 2.

final outcome. Moreover, many existing defenses primarily steer models toward more refusal-leaning behavior, often increasing over-refusal. 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, safety alignment in instruction-tuned models can be surprisingly shallow, with refusal often determined by a few early tokens and vulnerable to simple prefix manipulations (Qi et al., 2025; Liu et al., 2026; Zhao et al., 2025c; Yin et al., 2025). We extend this to reasoning models and show that refusal/compliance is already strongly readable at the start of the thinking trace and that current models make only limited use of thinking for safety deliberation.

Second, our findings relate to chain-of-thought faithfulness. In reasoning tasks, Boppana et al. (2026) and Cox et al. (2026) show that the final answer can often be decoded from hidden representations of the initial tokens for easier questions, suggesting that later reasoning tokens may partially rationalize an answer already decodable early on. We provide complementary evidence for safety: the apparently deliberative segments of the trace 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/compliance behavior is strongly readable at the first token, it is not 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 §4, 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. More optimistically, §3.2 suggests that the desired behavior is rare but present, opening the possibility of training objectives that reward real deliberation. We leave both as promising directions for future work.

Impact Statement

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.

Acknowledgements

We thank Luxi He and Boyi Wei for helpful discussions on early stages of this work, and Liam Fowl, Xingyu Zhu, and Ziyang Cai for providing feedback on drafts of this paper. NR acknowledges support by the Ezoe Memorial Recruit Foundation Scholarship and the Gordon Y.S. Wu Fellowship. AP acknowledges support by the Apple AI/ML Fellowship and Siebel Scholarship.

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 precot 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 additional 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 §3.1 show that even when individual segments of the thinking trace appear deliberative, the model often remains strongly biased toward its initial decision.

Table 4. Evaluation datasets. We select benchmarks that remain unsaturated for the models evaluated. Horizontal rule within ORR separates main and supplementary ORR sets. For WildJailbreak, we sample 2,000 test instances.

	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 5. Decoding settings for generation. All four models use nucleus sampling. Settings are taken from each model's published generation defaults. "–" indicates that the parameter is unset, in which case vLLM uses its default.

Model	**Temp.**	**Top-p**	**Top-k**
Qwen3-8B	0.6	0.95	20
Olmo-3-7B-Think	0.6	0.95	–
Phi-4-Reasoning (14B)	0.8	0.95	50
GPT-OSS-20B	1.0	1.0	–

B. Supplementary Details

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; the exact sampling parameters are listed in Table 5. All inference is performed with vLLM (Kwon et al., 2023) in BF16. For both the ASR and ORR splits, we cap the thinking trace at a budget of $16{,}384$ tokens. If a model emits its end-of-thinking marker (`</think>`, or the analysis-to-final channel switch for GPT-OSS-20B) within this budget, we use the natural completion as-is. Otherwise, we force-close the trace by appending the closure marker and generating a final response of up to $1{,}024$ additional tokens, so that every prompt yields a scorable answer rather than an unterminated trace. The context window is set to accommodate the prompt, the $16{,}384$-token thinking budget, and the final response. GPT-OSS-20B is generated in its native harmony channel format; before guardrail scoring, we strip all content outside the final channel. Under this rule, at most 0.88% of base-model completions reach the thinking budget without terminating and are force-closed.

First-Token Readout Features. For the trace-normalized Fisher discriminant of §2.2, we standardize each coordinate of the hidden state $\mathbf{h}_t$ to zero mean and unit variance over the pooled refusal/compliance sample at each position t before computing $J(t)$ to place all coordinates on a common scale. As $J(t)$ normalizes by the total within-class variance $\operatorname{tr} \mathbf{\Sigma}_W(t)$

rather than the full within-class covariance, this per-coordinate standardization affects the resulting value and is therefore applied consistently across all positions and models. The logistic probe of §2.2 analogously standardizes each feature dimension before projecting onto the top 100 principal components.

Balanced Accuracy via Youden's J Threshold. We compute the first-token probe's balanced accuracy (BAcc) by thresholding its refusal probability with Youden's J statistic. AUROC evaluates whether final-refusal examples receive higher probe scores than final-compliance examples, but it does not choose a binary decision threshold. 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. Youden's J is 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 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 §3.2, we prompt GPT-5.4, Gemini 3 Pro, and Sonnet 4.6 to assign stance labels to sentences indicating whether they lean toward compliance, refusal, or neither. See §C.3 for annotation evaluation.

Measuring Statistical Significance with p-Values. When we report whether a defense changes a quantity relative to its base model (e.g., first-token readability or the number of oscillations per trace in §4), we use a two-sided paired t-test on the per-cell base-to-trained differences. For a set of n model–defense cells, let $d_i = y_i^{\text{trained}} - y_i^{\text{base}}$ denote the change in the measured quantity for cell i, so that the cells provide paired observations. Under the null hypothesis that the defense induces no change on average ($\mathbb{E}[d_i] = 0$), the statistic

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}, \qquad \bar{d} = \frac{1}{n}\sum_i d_i, \qquad s_d = \sqrt{\frac{1}{n-1}\sum_i (d_i - \bar{d})^2},$$

follows a Student's t distribution with $n-1$ degrees of freedom, from which we obtain a two-sided p-value. We report a change as significant at level $\alpha = 0.05$.

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 context length of 24,576 tokens (16,384 for GPT-OSS-20B), following the same thinking-budget and force-close rule as our other evaluations.

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.

Prompt Template for Thinking Trace Annotation

```
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"}}
```

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 α=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. Depending on the method, the 8B/14B models are trained on one to five GPUs: the LoRA-based recipes (R1-ACT, STAIR, ThinkSafe) use one or two GPUs, and the full-parameter SFT recipes (e.g., STAR-1 and SafeKey) use two to five. GPT-OSS-20B uses DeepSpeed ZeRO-3 throughout, and all runs use 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 1×10^{-5}, cosine schedule with 5% linear warmup, effective batch size 256–640 (per-GPU batch size 1 with 128 gradient-accumulation steps, across two to five GPUs), 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 1×10^{-5}, weight decay 10^{-4}, $\beta = (0.9, 0.95)$, cosine schedule with 5% warmup, effective batch size 256–640, 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; for Olmo-3-7B-Think we use 5 epochs with the auxiliary heads active during the final 2.

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 1×10^{-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 2×10^{-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 5×10^{-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 1×10^{-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 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. Supplementary Results

C.1. ASR/ORR Evaluation

Table 6 reports the full per-benchmark ASR and ORR results for the four primary models across widely-used ASR and ORR benchmarks. For each prompt we draw $M = 4$ independent rollouts under each model's default sampling settings (Table 5) and evaluation setting used throughout the paper (§2.1). The main paper ASR evaluation uses WildJailbreak and FORTRESS,

Table 6. Base model ASR and ORR per benchmark. Lower is better for both. Values are success rates averaged over the four rollouts under the fractional four-guardrail vote; ± denotes a 95% bootstrap CI over the benchmark's prompts. Safest model per benchmark in bold. "Overall" is pooled over all prompts in the block.

	Benchmark	Qwen3-8B	Olmo-3-7B-Think	Phi-4-Reasoning	GPT-OSS-20B
ASR ↓	AdvBench	11.9 ±1.5	**3.4** ±0.9	3.7 ±0.9	4.1 ±0.9
	HarmBench	57.2 ±3.8	49.4 ±3.9	**8.9** ±1.6	12.8 ±1.6
	StrongREJECT	30.1 ±2.9	17.3 ±2.6	11.1 ±1.5	**10.4** ±1.4
	SORRY-Bench	51.5 ±3.7	36.4 ±3.5	30.5 ±3.6	**21.7** ±2.6
	JailbreakBench	29.8 ±6.4	15.1 ±5.0	11.9 ±4.7	**7.6** ±3.0
	WildJailbreak	85.7 ±1.2	68.0 ±1.5	47.6 ±1.8	**23.0** ±1.2
	FORTRESS	91.2 ±1.5	88.7 ±1.7	51.1 ±3.1	**44.9** ±3.1
	HEx-PHI	32.0 ±3.6	19.6 ±3.2	15.1 ±3.0	**8.6** ±1.6
	Overall	63.6 ±1.2	50.4 ±1.2	32.5 ±1.1	**20.1** ±0.8
ORR ↓	OR-Bench	**6.4** ±0.7	23.8 ±1.4	45.8 ±1.9	58.1 ±1.6
	FalseReject	**7.9** ±1.0	16.5 ±1.5	43.0 ±2.2	42.9 ±2.1
	CoCoNot	**0.1** ±0.2	0.6 ±0.3	1.5 ±0.7	3.0 ±1.2
	XSTest	**1.1** ±0.7	3.9 ±1.5	11.1 ±3.0	7.9 ±2.7
	Overall	**5.8** ±0.5	16.6 ±0.8	36.6 ±1.2	41.7 ±1.2

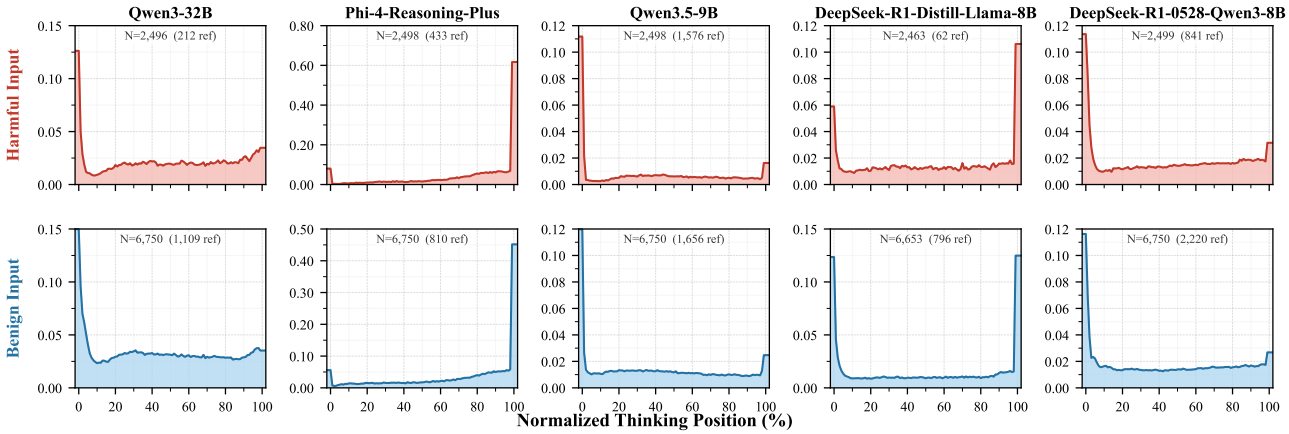

Figure 7. Fisher discriminant plot for supplementary thinking models. All five models reproduce the same qualitative valley signature observed in the four primary models, with high pre-decoding separability, a sustained drop through the middle of the trace, and re-crystallization at the end.

which exhibit well-separated ASR across all four models (12–97%). On most of the remaining ASR benchmarks the more strongly aligned models saturate near the floor, with both GPT-OSS-20B and Phi-4-Reasoning complying with under 2% of AdvBench and StrongREJECT prompts, which leaves little range for separating models or measuring the effect of interventions. For ORR, we use OR-Bench, FalseReject, and CoCoNot as the over-refusal pool, as they are widely-used standard benchmarks for measuring over-refusal; we additionally report XSTest here for completeness.

C.2. Fisher Discriminants

We present Fisher discriminant plots for additional thinking models in Figure 7. Qwen3-32B is the 32B-parameter variant of Qwen3 and uses the same thinking-mode chat template as our primary Qwen3-8B model. Qwen3.5-9B is a 9B-parameter model from the more recent Qwen3.5 generation, testing whether the effect persists across successive Qwen releases. 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.

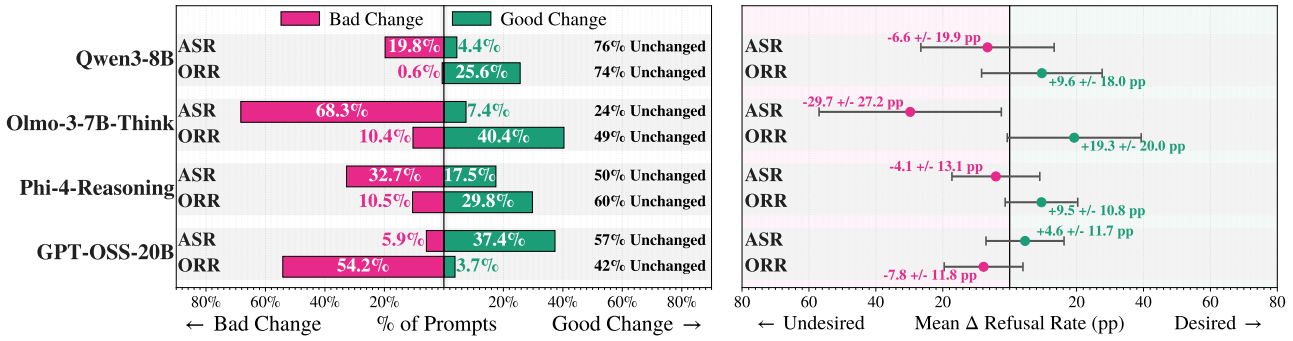

Figure 8. More permissive measures of the no-think to think change. Left: sign-only directional changes. Any nonzero movement toward the desired refusal/compliance behavior is counted as a desired-direction change. Right: mean signed change in refusal rate with standard deviation over inputs. Positive values indicate more refusal on ASR inputs and less refusal on ORR inputs.

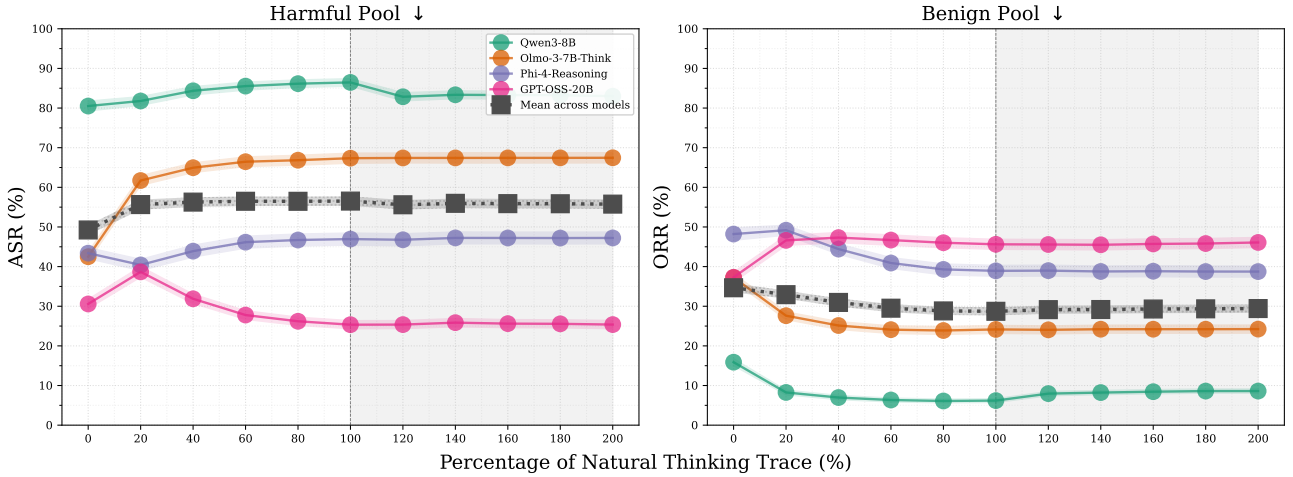

Figure 9. ASR on the harmful pool (left) and ORR on the benign pool (right) for each thinking budget B. For $B < 100\%$, the thinking trace is truncated to the given fraction of its natural length before the model is forced to produce a final answer. For $B > 100\%$ (shaded region), the model is forced to continue thinking past its natural stopping point by masking the end-of-thinking marker and sampling further. Lower is better for both metrics. Results use the full $K{=}32$ pools of 2,500 harmful and 2,885 benign prompts. Across most models, increasing the thinking budget tends to raise compliance (higher ASR and lower ORR), while forcing thinking beyond the natural stopping point yields little additional change.

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.

C.3. Effect of Thinking

More Permissive No-Think-to-Think Change Measures. The majority-label flip measure in §3.1 captures whether thinking changes the dominant input-level decision, but does not credit small shifts in the desired direction unless they cross the majority threshold. We therefore consider two more permissive measures of the no-think to think change. For each input, let q_{NT} and q_{T} respectively denote the estimated refusal rates under the no-think and think conditions. We orient the difference so that positive values always indicate movement toward the desired behavior:

$$\Delta_{\text{desired}} = \begin{cases} q_{\mathrm{T}} - q_{\mathrm{NT}}, & \text{ASR inputs,} \\ q_{\mathrm{NT}} - q_{\mathrm{T}}, & \text{ORR inputs.} \end{cases}$$

Table 7. Per-annotator statistics. We use GPT-5.4, Gemini 3 Pro, and Sonnet 4.6 as the main annotators, and Opus 4.7 as a reference auditor. Mean pairwise κ averages an annotator's Cohen κ with the other three models.

	Main Annotators			**Reference Annotator**
	GPT-5.4	Gemini 3 Pro	Sonnet 4.6	Opus 4.7
Traces labeled (of 500)	500	496	477	453
Refusal-leaning (-1, %)	29.8	33.7	38.9	36.4
Neutral (0, %)	43.6	35.2	32.9	30.3
Compliance-leaning ($+1$, %)	26.6	31.1	28.2	33.3
Mean pairwise Cohen κ	0.632	0.659	0.646	0.663

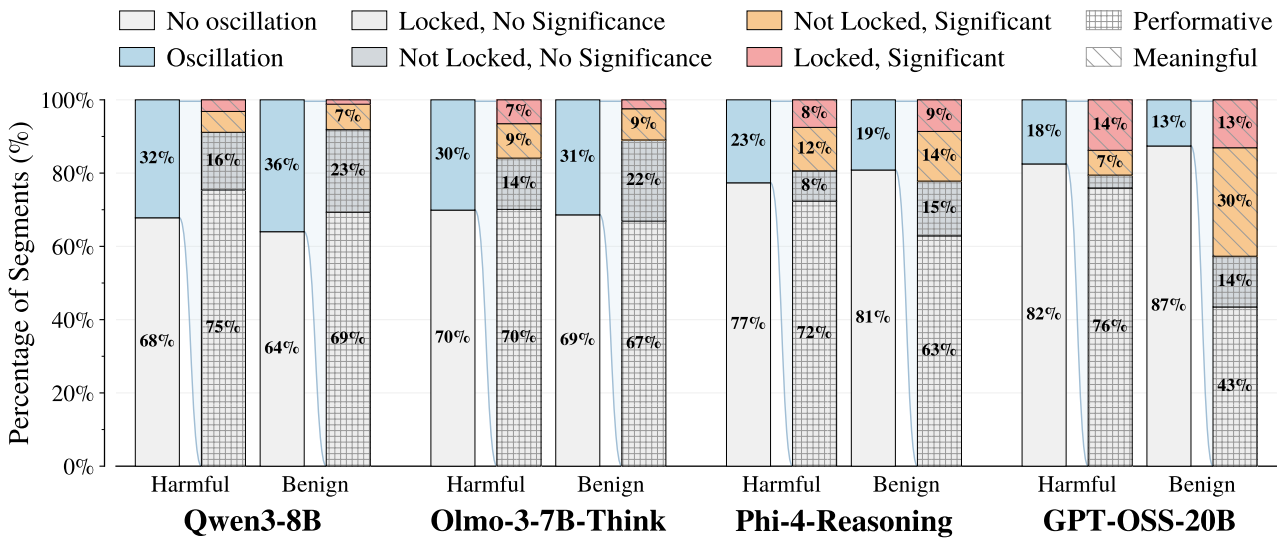

Figure 10. Per-split breakdown of thinking trace oscillations and their causal effect. For each model and split, the overview bar shows the percentage of $M{=}4$ thinking rollouts exhibiting at least one stance oscillation, and the exploded bar decomposes those oscillation segments into four categories along the Locked / Not Locked axis and the Performative / Meaningful axis. In both splits, most oscillation segments are performative: the reasoning wavers but the $K{=}100$ resampled final-response distribution does not significantly move.

The first measure binarizes this signed shift and counts an input as moving in the desired direction whenever $\Delta_{\text{desired}} > 0$ and in the undesired direction whenever $\Delta_{\text{desired}} < 0$; this is the most lenient directional criterion, as any nonzero improvement is counted. The second measure examines the magnitude of the shift and reports the mean and standard deviation of Δ_{desired} over inputs.

See Figure 8 for the results. Phi-4-Reasoning shows a mild exception, but both measures largely support the same qualitative conclusion as the majority-label flip analysis and corroborate the main takeaway.

Extended Thinking and ASR/ORR. In §3.1, we found that continuation variance given a fixed prefix is very low and that enabling thinking does not improve model behavior overall. Here, we ask whether the ASR–ORR tradeoff changes when we control the amount of thinking more directly. We evaluate both limited thinking budgets and extended thinking beyond the model's natural stopping point. For the latter, we mask the end-of-thinking token and continue sampling from the remaining distribution until the extra budget is exhausted.

See Figure 9 for the results. Although GPT-OSS-20B improves ASR with additional thinking, this tends to come at the cost of increasing ORR. While the remaining models show the opposite failure mode, we see a trend where more thinking increases compliance, consistent with the no-think versus think results in Figure 4. Furthermore, forcing models to think past their natural stopping point does not meaningfully change behavior. Overall, we find weak evidence that more thinking improves the ASR–ORR tradeoff.

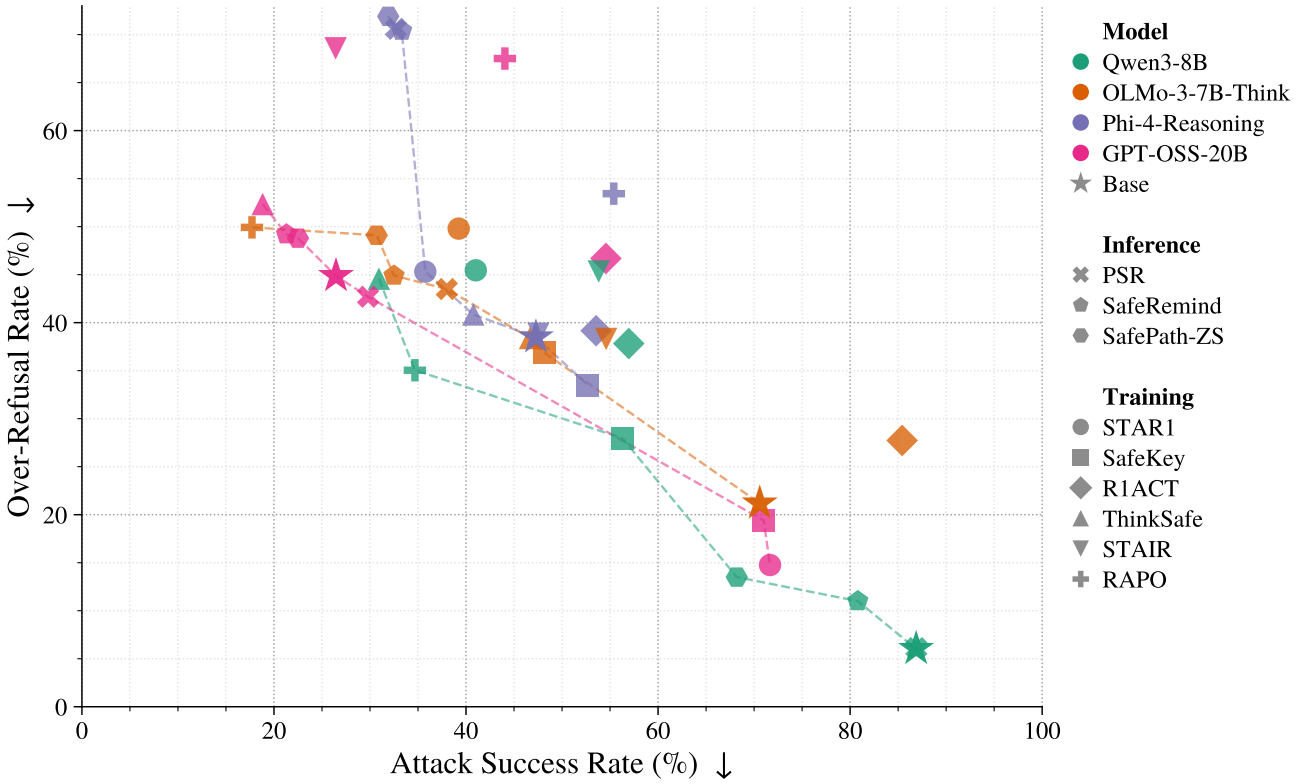

Figure 11. Pareto plot of ASR and ORR for every defense. Each point represents one base model paired with one defense. Pareto-optimal cells lie toward the bottom-left. Color denotes the base model, marker denotes the defense method, and star marks the undefended base model. Dashed lines trace each base model's Pareto frontier over its own defenses. Most defenses move along their base model's ASR–ORR tradeoff rather than shifting it toward the bottom-left.

Inter-Annotator Agreement (IAA) of Thinking Trace Annotators. We assess IAA among three frontier auditors: GPT-5.4, Gemini 3 Pro, and Claude Sonnet 4.6, whose per-segment majority vote provides the stance labels used in the analyses of §3.2 and §4. We sample N=500 thinking traces uniformly at random with 250 from base models and 250 from a defense subset, balanced across the four base models and two prompt splits and further balanced within stratum. To gauge whether Sonnet 4.6 is sufficient as the third main annotator, we additionally run Claude Opus 4.7 over the same traces as a quality reference auditor. We note that as Opus 4.7 declines to label a non-trivial fraction of traces (see below), we report statistics on 9,463 sentence-level segments from 436 traces that all four auditors labeled.

We report Fleiss' κ for three-way agreement and the mean of pairwise Cohen κ with the other auditors for each annotator. Because no gold labels exist for thinking trace stance, we also report each annotator's match rate against the per-segment majority-of-three over the main annotators (GPT-5.4, Gemini 3 Pro, Sonnet 4.6) a centrality measure. This proxy is reasonable in our setting, as agreement is substantial overall, full three-way disagreement is rare, and the annotators come from distinct model families.

We find substantial agreement among the three main auditors, with a Fleiss κ of 0.636 ± 0.031 overall, three-way concordance on 64.4% of segments, and full three-way disagreement on only 1.3%. Agreement is stable across strata (base traces 0.610, defense traces 0.672) and holds within every base model, ranging from 0.574 for Qwen3-8B and 0.581 for Olmo-3-7B-Think to 0.662 for Phi-4-Reasoning and 0.709 for GPT-OSS-20B. We also note that agreement is essentially unchanged on the larger set of 469 traces (10,193 segments) jointly labeled by the three main auditors (Fleiss κ=0.632 ± 0.029), indicating that restricting to the four-way common set does not inflate agreement. We further report per-annotator statistics in Table 7. Observe that each annotator's mean pairwise Cohen κ with the others is comparable (0.632–0.663), and label marginals differ only mildly. GPT-5.4 most often assigns the neutral label, Sonnet 4.6 is the most refusal-leaning, and Gemini 3 Pro most often assigns the compliance-leaning label.

As mentioned, we use Opus 4.7 only as a reference for whether Sonnet 4.6 is an adequate third annotator, and two

observations support this choice. First, Opus 4.7 declines to label more traces than Sonnet 4.6 (9.4% vs. 4.6% of the 500 traces, almost entirely on harmful prompts). Second, on the common set, Sonnet 4.6 agrees with the other auditors about as well as Opus 4.7 does (mean pairwise κ 0.646 vs. 0.663), and replacing Sonnet with Opus as the third annotator changes the three-way Fleiss κ by only 0.017 ($0.636 \rightarrow 0.653$). We therefore adopt Sonnet 4.6 as the third main annotator and retain Opus 4.7 as a reference. For all analyses, we use the majority-of-three label over the main annotators when a majority exists.

C.4. Per-Split Breakdown of Thinking Trace Effect

Figure 5 pools harmful (ASR) and benign (ORR) prompts when reporting the per-rollout oscillation rate and the per-segment Locked $\times$ Significant decomposition. Figure 10 reports the same two quantities separately for the two prompt pools. For each model, the left group corresponds to harmful (ASR) prompts and the right group to benign (ORR) prompts. Within each group, the overview bar gives the fraction of the $M{=}4$ rollouts that contain at least one oscillation, and the adjacent exploded bar decomposes that pool's oscillation segments into the four Locked $\times$ {Performative, Meaningful} categories defined in §3.2 and App. B. The qualitative conclusion of §3.2 holds within each split: across both harmful and benign prompts, the large majority of oscillation segments are performative, with only a small statistically significant (Meaningful) fraction crossing the refuse/comply boundary.

C.5. Defense Evaluation

Figure 11 plots the ASR and ORR of every defense from Table 2 on the ASR–ORR plane. Note that the ideal region is the bottom-left corner of low ASR and ORR. Observe that no defended cell improves on its base model in both metrics at once across all nine defenses applied to the four base models; every cell that lowers ASR does so by raising ORR, a smaller number instead trade higher ASR for lower ORR, and a handful are worse than the base on both axes. This pattern holds for both inference-time and training-based defenses, and each base model's Pareto frontier is traced largely by the undefended model together with cells that shift along the same tradeoff, rather than by defenses that push the frontier toward the bottom-left. The conclusion of §4 clearly holds in that the defenses we evaluate reallocate refusal behavior between harmful and benign prompts rather than improving the underlying ASR–ORR tradeoff.