

The Geometry of Safety Drifts: Step-level Refusal-Subspace Monitoring for Reasoning Large Language Models

Liming Lu¹, Anan Du², Fanzhen Liu³, Hechang Chen⁴, Shuchao Pang^{1*}

¹Nanjing University of Science and Technology ²Nanjing University of Industry Technology

³Macquarie University ⁴Jilin University

{luliming, pangshuchao}@njust.edu.cn, anan.du@niit.edu.cn, fanzhen.liu@mq.edu.au

chenhc@jlu.edu.cn

Abstract

Reasoning large language models that produce explicit chains of thought are a new jailbreak surface: an adversary who shapes intermediate reasoning can hijack safety deliberation and elicit harmful answers the same model would refuse in a single forward pass. Defenses inherited from chat-style alignment monitor only the final response and intervene after the model has already committed to an unsafe trajectory. We argue that defending reasoning models requires looking inside the thinking phase itself, and we show that the refusal direction, previously characterized as a single linear direction in chat models, in fact evolves along the reasoning trajectory: sharp at onset, slowly drifting during long deliberation, and on hijacked traces collapsing at an identifiable step well before any harmful token is emitted. We propose RSMON, a step-level refusal-subspace monitor that tracks the projection of intermediate hidden states onto a calibrated refusal subspace and triggers a minimally invasive geometric intervention once drift exceeds a learned threshold. The monitor is purely inference-time, requires no retraining, and is compatible with any open-weight reasoning model. On three open reasoning models, five attack families, and four utility benchmarks, RSMON cuts mean attack success from 80.1% to 19.2%, a $4.2\times$ reduction, while keeping MATH-500, GPQA-Diamond, and HumanEval pass@1 within 0.7 absolute points and MT-Bench within 0.08 points of the undefended baseline. The same monitor produces a step-resolved diagnostic that pinpoints when and in which layer safety deliberation breaks down. Our code is available at <https://github.com/llmnjust-afk/RsMon>.

1 Introduction

Reasoning large language models (RLLMs) such as DeepSeek-R1 (Guo et al., 2025), the OpenAI

o-series (OpenAI, 2024), and the Qwen3 thinking variants (Yang et al., 2025) extend standard autoregressive generation with an explicit, lengthy chain of thought before producing an answer. This thinking phase delivers state-of-the-art performance on mathematics, code, and scientific reasoning (Chen et al., 2026; Shojaei et al., 2025), and the same paradigm has been argued to also enable safer behavior through deliberation (Guan et al., 2024). The thinking phase is also, by construction, an unprecedented enlargement of the model’s surface area: every intermediate token is a step in which the model can be biased, distracted, or persuaded. The community has only recently begun to map the resulting safety landscape (Wang et al., 2025a; Mercer et al., 2025).

Recent red-teaming work has weaponized exactly this surface. Kuo et al. (2025) introduce H-CoT, which jailbreaks o1/o3, DeepSeek-R1, and Gemini 2.0 Flash Thinking by injecting fragments that hijack the model’s own safety reasoning. Zhao et al. (2025a) generalize the phenomenon under the name *Chain-of-Thought Hijacking* and show that longer reasoning, far from improving safety, can monotonically erode it on harmful queries. Yang (2026) further show that thinking models can be broken by multi-stream perturbations that interfere with their concurrent task representations, while Ying et al. (2025) demonstrate that multi-turn reasoning conversation amplifies attack success. Arcuschin et al. (2025) further show that chain-of-thought reasoning is not always faithful to the model’s actual decision process.

Existing defenses, however, were designed for chat-style large language models. Input-side filters (Hu et al., 2024; Jain et al., 2023), output-side guardrails (Zou et al., 2024; Inan et al., 2023), decoding-time interventions (Xu et al., 2024a; Li et al., 2024c), certified detectors (Kumar et al., 2023; Robey et al., 2023), layer-level edits (Zhao et al., 2024), and in-context demonstrations (Wei

*Corresponding author.

et al., 2026; Zhang et al., 2024) all operate on a single forward pass: they treat the model’s safety decision as a one-shot event. In a reasoning model that may emit thousands of intermediate tokens before committing to an answer, such defenses inherit a structural blind spot: the model can drift into an unsafe deliberation step by step, and the defense will only react after the unsafe answer has begun. Survey work has identified this gap as one of the central open problems for reasoning-model safety (Wang et al., 2025a; Dong et al., 2025; Xu et al., 2024b).

We address this blind spot by asking a mechanistic question. Arditi et al. (2024) showed that in chat large language models refusal behavior is mediated by a single linear direction in activation space, and Wang et al. (2025b); Pan et al. (2025) extended the picture to cross-lingual and multi-dimensional refusal subspaces. Alagharu et al. (2026) and Lee et al. (2025) further showed that refusal can be conditionally programmed by steering along these directions, and Zhang and Sun (2026) showed that the same geometry can be exploited adversarially. All of this prior evidence, however, was collected in single-step chat settings; the temporal structure of refusal inside a long reasoning trace has not been characterized. We ask: what happens to the refusal geometry along the reasoning trajectory?

Overall, our contributions are summarized as follows: **(I) Empirical phenomenon.** We characterize, for the first time, how the refusal direction behaves *step by step* during the thinking phase of three open RLLMs: DeepSeek-R1-Distill-Qwen-7B, Qwen3-8B-Thinking, and QwQ-32B-Preview. We find a consistent three-regime structure consisting of a sharp *refusal onset*, a slow *utility drift*, and on attacked inputs an abrupt *safety collapse* that precedes any harmful token by several steps. **(II) Diagnostic tool.** We turn this phenomenon into a step-level diagnostic, the *Refusal-Subspace Drift Curve*, that visualizes for any reasoning trace where and how safety deliberation breaks down. The curve gives a model-agnostic answer to the question “at which step did the model give up refusing?” **(III) Defense.** We propose RSMON, a step-level *Refusal-Subspace Monitor* with a minimally invasive geometric intervention. RSMON projects intermediate residual-stream activations onto a calibrated refusal subspace, detects when the projection departs from a benign-trajectory envelope, and applies one of three increasingly invasive interventions: subspace re-projection, controlled rollback

of the thinking trace, or early-termination with a safe refusal. The whole pipeline is purely inference-time, requires no fine-tuning, and works on any open-weight RLLM whose residual stream is accessible. **(IV) Evaluation.** We evaluate against five attack families on three benchmarks while preserving reasoning utility on four capability benchmarks, follow the stability protocol of Larsen (2025) to avoid single-shot bias, and run a worst-case adaptive evaluation in which the adversary explicitly optimizes against the monitor. RSMON dominates every training-free baseline on the joint safety–utility plane and matches the fine-tuned topline REALSAFE-R1 (Zhang et al., 2025b) without any retraining.

Key Differences from Existing Approaches. Our contribution is positioned against, and explicitly distinct from, the most closely related concurrent work. REALSAFE-R1 (Zhang et al., 2025b) hardens DeepSeek-R1 by fine-tuning, whereas our method is training-free and complementary. Circuit Breakers (Zou et al., 2024) intervene on the final representation, whereas we intervene at intermediate reasoning steps. Alagharu et al. (2026) and Lee et al. (2025) steer refusal in chat models, but we are the first to characterize and exploit the temporal structure of refusal in long reasoning traces. An overview of the full pipeline is given in Figure 1.

2 Related Work

Jailbreak attacks on standard large language models. Optimization-based attacks such as GCG and its successors construct adversarial suffixes that maximize the likelihood of an affirmative prefix (Li et al., 2024b; Mu et al., 2026). Black-box attacks instead leverage natural-language persuasion (Zeng et al., 2024), contextual priming across turns (Miao et al., 2025; Du et al., 2025; Ying et al., 2025), ASCII-art obfuscation (Jiang et al., 2024), nested scene templates (Ding et al., 2024), intent manipulation against content moderators (Zhuang et al., 2025), in-the-wild prompt libraries (Shen et al., 2024), or simple adaptive search using log-prob feedback (Andriushchenko et al., 2025; Chao et al., 2025). Comprehensive surveys agree that no deployed alignment technique is robust against all of these families (Yi et al., 2024; Xu et al., 2024b; Shayegani et al., 2023; Dong et al., 2025), and recent stability analysis shows that single-shot evaluations systematically underestimate true attack risk (Larsen, 2025).

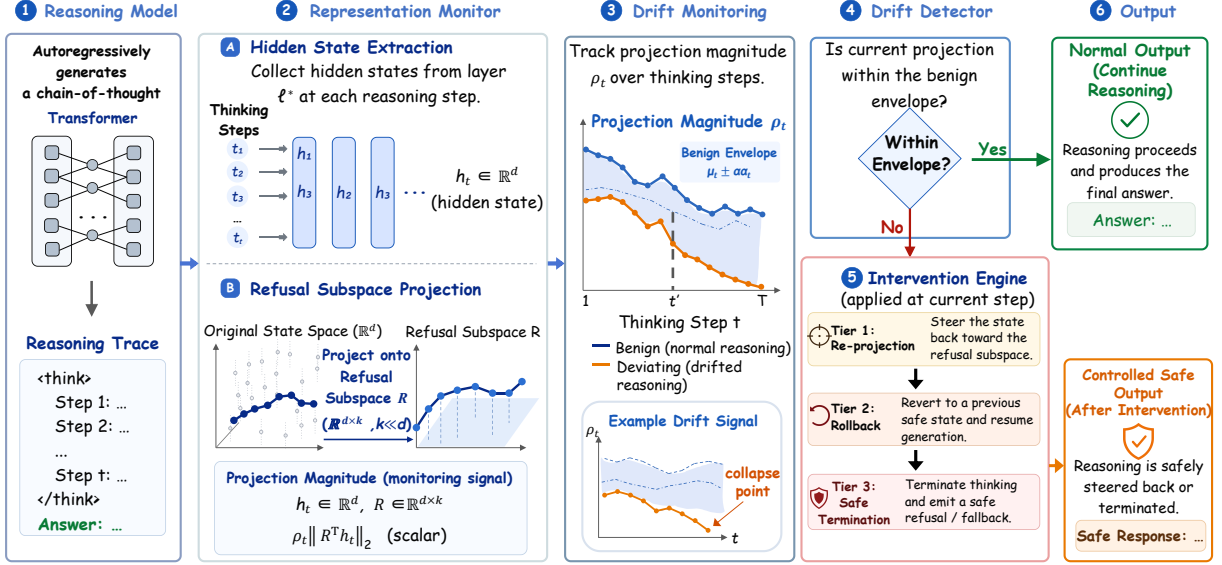


Figure 1: Overview of RSMON. Given a reasoning model that generates an explicit chain-of-thought trajectory, RSMON monitors the evolution of intermediate hidden states during the thinking phase. At each reasoning step, the residual-stream activation from a calibrated layer is extracted and projected onto a learned refusal subspace. The resulting projection magnitude is tracked as a refusal-drift signal and compared against a benign envelope to identify safety deviations. When a drift is detected, RSMON applies a progressively stronger intervention strategy, including refusal-subspace re-projection, reasoning rollback, and safe termination, to steer the model back to a safe output. The entire procedure is training-free and operates at inference time.

Jailbreaking reasoning models. The reasoning-model regime is qualitatively different. Kuo et al. (2025) show that the safety-as-reasoning mechanism used by frontier RLLMs can be hijacked by injecting fragments that mimic the model’s safety deliberation. Zhao et al. (2025a) generalize this into a broader chain-of-thought hijacking phenomenon and demonstrate that longer reasoning can monotonically degrade safety. Yang (2026) introduce multi-stream perturbation attacks that interfere with concurrent task representations inside thinking models, and Ying et al. (2025) show that reasoning-augmented multi-turn conversation further amplifies attack success on aligned chat models. Nie et al. (2026) document a related coupling effect in which a fixed output budget couples reasoning and answer tokens, indirectly degrading safety. A concurrent survey (Wang et al., 2025a) catalogues these phenomena and identifies trajectory-level defenses as an open problem; our work directly targets this gap.

Defenses for reasoning models. REALSAFE-R1 (Zhang et al., 2025b) retrains DeepSeek-R1 with safety-aware reasoning supervision and demonstrates that safety can be improved without sacrificing reasoning quality. Deliberative alignment (Guan et al., 2024) similarly trains the model

to reason explicitly about safety policies before answering. Both are closely related to our approach, but require costly retraining for each model release and, as single-step alignment interventions, provide little insight into when safety failures emerge. Our method is orthogonal: it is training-free, inference-time, and produces a mechanistic diagnostic in addition to a defense.

3 Background and Threat Model

3.1 Reasoning Large Language Models

We focus on autoregressive transformer reasoning models that emit a structured trace of the form

$$[\text{<think>} t_1, t_2, \dots, t_T \text{</think>} a_1, a_2, \dots, a_M],$$

where $t_{1:T}$ is the *thinking phase* and $a_{1:M}$ is the user-visible *answer phase*. We assume access to the residual stream at every transformer layer for every token, but no access to model weights for fine-tuning. This setting is realized by, for example, DeepSeek-R1-Distill (Guo et al., 2025), Qwen3-Thinking (Yang et al., 2025), and QwQ-Preview (Qwen Team, 2024). Recent analysis of these architectures confirms that the explicit thinking phase consistently increases capability on multi-step reasoning benchmarks but also introduces new failure modes specific to extended

deliberation (Shojaee et al., 2025; Mercer et al., 2025; Nie et al., 2026).

3.2 The Refusal Direction

Following Arditi et al. (2024), given a chat large language model and paired harmful/harmless prompts, the difference in mean residual-stream activations defines a refusal direction $\mathbf{r} \in \mathbb{R}^d$ such that projecting hidden states onto $\mathbf{r}$ reliably separates refusing from complying behavior in a single forward pass. Pan et al. (2025) generalize $\mathbf{r}$ to a k -dimensional refusal subspace $\mathcal{R} \subset \mathbb{R}^d$ spanned by orthogonal directions associated with distinct harm categories, and Wang et al. (2025b); Alagharu et al. (2026) establish that the same geometric structure recurs across languages and across categorical refusal tokens. We adopt the multi-dimensional view throughout, and we ablate the choice of k in Section B. Critically, all of this prior evidence comes from single-step chat settings; the present paper asks whether the same geometric object survives, and how it evolves, along a long thinking trajectory.

3.3 Threat Model

The adversary submits a single user prompt x to the reasoning model and observes the full output (thinking plus answer). The adversary may use white-box gradient access for optimization-based attacks (Li et al., 2024b; Mu et al., 2026) or black-box query-only access for natural-language attacks (Zeng et al., 2024; Kuo et al., 2025; Zhao et al., 2025a; Andriushchenko et al., 2025; Miao et al., 2025; Ying et al., 2025; Yang, 2026). The defender controls the model’s inference pipeline: at every generation step the defender may read the residual stream and may write a bounded perturbation to it. The defender may not retrain the model, may not modify the system prompt visible to the user, and must preserve reasoning utility on benign inputs as measured by the four capability benchmarks of Section 5. We additionally consider an *adaptive* adversary who knows the defense and explicitly optimizes the attack to keep the drift curve inside the benign envelope; results for this worst case are reported in Section 6.

4 Method: RSMON

RSMON is a three-component inference-time procedure that converts the drift curve into a defense. The overall data flow is summarized in Figure 1, and the underlying geometric intuition is given in

Figure 2. Each component is independently configurable and can be replaced by a stronger drop-in variant without changing the rest of the pipeline.

4.1 Refusal Drift in Reasoning Models

Let $h_t^{(\ell)} \in \mathbb{R}^d$ denote the residual-stream activation at layer ℓ and thinking step t . Given a calibrated refusal subspace $\mathcal{R}$ with orthonormal basis $\mathbf{R} \in \mathbb{R}^{d \times k}$, we define the *step-level refusal projection*

$$\rho_t^{(\ell)} = \|\mathbf{R}^\top h_t^{(\ell)}\|_2, \quad (1)$$

and call the sequence $\{\rho_t^{(\ell)}\}_{t=1}^T$ the *Refusal-Subspace Drift Curve* at layer ℓ . We empirically observe three qualitatively distinct regimes of this curve. On any harmful query, $\rho_t^{(\ell)}$ rises sharply in the first few thinking steps as the model triggers its safety deliberation (**onset**). During long benign reasoning, $\rho_t^{(\ell)}$ slowly decays as the residual stream is dominated by task-relevant features unrelated to refusal (**utility drift**). On hijacked harmful queries, the curve exhibits an abrupt downward step at an identifiable point t^* that precedes the emission of any harmful answer token by several steps (**safety collapse**). Figure 3 visualizes the three regimes on DeepSeek-R1-Distill-Qwen-7B. Across all three open reasoning models we study, the benign and harmful regimes occupy distinct, stable bands of $\rho_t^{(\ell^*)}$, while hijacked traces undergo an abrupt downward step at a per-trace collapse step t^* that we recover automatically from the curve.

4.2 Calibration of the Refusal Subspace

Following Arditi et al. (2024) and Pan et al. (2025), we estimate $\mathbf{R}$ from a small held-out set of paired harmful/harmless prompts $\{(x_i^-, x_i^+)\}_{i=1}^N$. For each layer ℓ^* chosen by linear-probe accuracy on a validation split (Figure 7), we compute the per-prompt mean activation in the first τ_0 thinking steps and take the top- k principal directions of the difference matrix to form $\mathbf{R} \in \mathbb{R}^{d \times k}$. We use $N = 512$, $\tau_0 = 8$, and $k \in \{1, 2, 4, 8\}$ (ablated in Section B). The calibration uses only freely available harmful-prompt corpora, specifically the held-out splits of HarmBench (Mazeika et al., 2024) and SALAD-Bench (Li et al., 2024a), and no model weights are ever touched.

4.3 Benign Envelope and Drift Detector

For each layer ℓ^* we collect drift curves on a benign-reasoning corpus drawn from

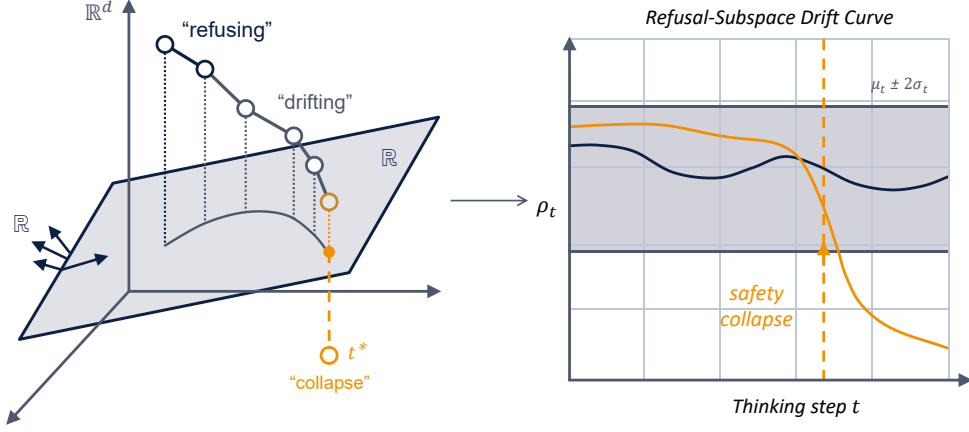


Figure 2: Geometric intuition for RSMON. *Left*: hidden states h_t at successive thinking steps trace a path in $\mathbb{R}^d$; refusal is mediated by a k -dimensional subspace $\mathcal{R}$ (oblique plane), and the projection of h_t onto $\mathcal{R}$ measures how strongly the current deliberation lies in the refusal regime. *Right*: the resulting drift curve ρ_t for benign reasoning stays inside a calibrated envelope, while a hijacked trajectory undergoes an abrupt collapse below the envelope at step t^* ; this is the firing signal exploited by the defense.

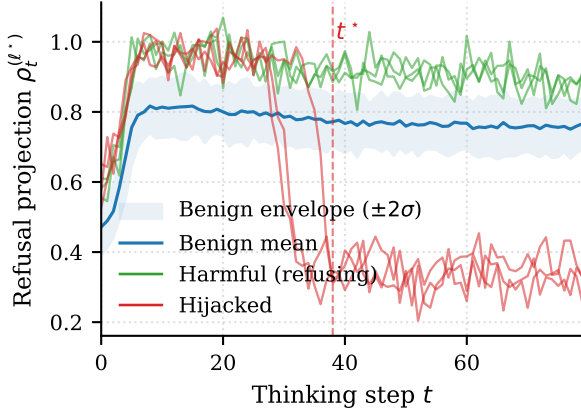


Figure 3: The Refusal-Subspace Drift Curve $\rho_t^{(\ell^*)}$ on three regimes (DEEPSEEK-R1-DISTILL-QWEN-7B). The benign envelope (shaded band, $\pm 2\sigma$) calibrated on benign reasoning corpora cleanly separates the benign and harmful-refusing trajectories. On hijacked traces, ρ undergoes an abrupt collapse at t^* (dashed line), preceding any harmful answer token by several steps.

MATH (Hendrycks et al., 2021), GPQA (Rein et al., 2023), and a held-out set of instruction-following prompts. We then fit a per-step lower envelope $\rho_t^{(\ell^*)} = \mu_t - \alpha\sigma_t$, where μ_t and σ_t are the empirical mean and standard deviation of $\rho_t^{(\ell^*)}$ across benign traces at thinking step t . The detector fires at step t if

$$\rho_t^{(\ell^*)} < \underline{\rho}_t^{(\ell^*)} \quad \text{for } \Delta \text{ consecutive steps,} \quad (2)$$

with $\Delta \in \{1, 2, 3\}$. The two thresholds α and Δ jointly control the false-positive rate on benign reasoning; we tune them on a validation split to keep benign utility loss below a fixed budget of 1% on

MATH. This corresponds to a fixed-budget calibration that mirrors the operating-point selection used by SafeDecoding (Xu et al., 2024a) and Gradient Cuff (Hu et al., 2024), but applied per thinking step rather than once per query.

4.4 Three-Tier Intervention

When the detector fires at step $t^\dagger$, RSMON applies the least invasive intervention that restores the projection into the benign envelope. The first tier is *re-projection*: replace $h_{t^\dagger}^{(\ell^*)}$ with $h_{t^\dagger}^{(\ell^*)} + \beta \mathbf{R} \mathbf{R}^\top \bar{h}_{t^\dagger}$, where $\bar{h}_{t^\dagger}$ is the benign-mean activation at step $t^\dagger$ and β is the smallest scalar that pushes the projection above $\underline{\rho}_t^{(\ell^*)}$. This is a single bounded edit to the residual stream and preserves task-relevant content in the directions orthogonal to $\mathcal{R}$. The second tier is *rollback*: if a second detector firing occurs within $W = 16$ steps, the thinking trace is rolled back to step $t^\dagger - \kappa$ and resumed with a refusal-priming token (`Wait, ...`). This reuses the model’s own deliberation rather than overriding it, and is in spirit similar to the auto-regressive self-alignment of Li et al. (2024c). The third tier is *hard refusal*: if the rollback is itself flagged, the thinking phase is terminated and a templated safe refusal is emitted as the answer. Tier escalation is monotone, deterministic, and bounded; in the worst case RSMON emits a hard refusal at $O(T)$ additional compute.

RSMON satisfies five desirable properties. *Training-free*: no parameters of the base model are updated. *Architecture-agnostic*: requires only read/write access to the residual stream, which every open-weight transformer reasoning model in

our study supports. *Bounded*: inference overhead is $O(kd)$ per generated token plus an $O(1)$ detector check. *Interpretable*: every firing records the step at which drift exceeded the envelope and the layer that detected it. *Compatible*: stacks with existing chat-style defenses such as Circuit Breakers, Gradient Cuff, and SafeDecoding, which we treat as additional baselines in Section 5.

5 Experimental Setup

5.1 Models

We evaluate on three open-weight reasoning models released within the last six months prior to submission: DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025), Qwen3-8B-Thinking (Yang et al., 2025), and QwQ-32B-Preview (Qwen Team, 2024). All three expose an explicit thinking phase and accessible residual stream, span two distinct model families (DeepSeek/Qwen), and cover the 7B–32B parameter range that dominates current open reasoning-model deployment. We additionally report cross-model transfer of the calibrated refusal subspace in Section B. Older non-thinking models are used only as historical references in the related-work discussion and are not part of the evaluation matrix.

5.2 Attack Suite

We use five attack families spanning the current state of the art. *Chain-of-Thought Hijacking* (Zhao et al., 2025a) is the primary reasoning-model attack and constitutes our headline column. *H-CoT* (Kuo et al., 2025) hijacks the safety reasoning fragment directly. *Faster-GCG* (Li et al., 2024b) is the strongest current optimization-based adversarial-suffix attack. *Persuasion* (Zeng et al., 2024) is the strongest natural-language social-engineering attack. *Simple Adaptive Attack* (Andriushchenko et al., 2025) is a query-only black-box attack using logprob feedback that has been shown to defeat many chat-side defenses. For each attack we use the harmful-behavior split of HarmBench (Mazeika et al., 2024) (159 behaviors) and report attack success rate (ASR) judged by the HarmBench classifier. We additionally evaluate on SALAD-Bench (Li et al., 2024a) and on the renewable JB-Distill benchmark (Zhang et al., 2025a) to mitigate test-set leakage and evaluation drift, and we cross-check ASR with the R-Judge protocol (Yuan et al., 2024) on a random subset of 300 traces.

5.3 Defense Baselines

We compare RSMON against eight baselines spanning every major family of chat-style defenses. *No defense* is the base reasoning model. *System prompt* is a strong safety system prompt patterned after the goal-prioritization template of Zhang et al. (2024). *Gradient Cuff* (Hu et al., 2024) is an input-side detector. *Circuit Breakers* (Zou et al., 2024) is a representation-engineering defense on the final hidden state, adapted to the answer phase only as in the original paper. *Layer-Specific Editing* (Zhao et al., 2024) is an offline edit applied to selected layers. *In-Context Demonstrations* (Wei et al., 2026) is a few-shot safety guard. *Proactive Defense* (Zhao et al., 2025b) is a multi-turn drift defense, applied to the thinking trace as a strong adaptation. REALSAFE-R1 (Zhang et al., 2025b) is the fine-tuned safety variant of DeepSeek-R1; we report it as a training-based topline reference, not as a like-for-like baseline, since it requires retraining.

5.4 Utility Benchmarks

To verify that RSMON does not damage reasoning quality, we report MATH-500 accuracy (Hendrycks et al., 2021), GPQA-Diamond accuracy (Rein et al., 2023), HumanEval pass@1 (Chen et al., 2021), and MT-Bench score (Zheng et al., 2023) for instruction following. These four benchmarks were chosen to cover mathematical reasoning, scientific reasoning, code generation, and open-ended instruction following; the inclusion of MT-Bench specifically guards against the failure mode in which a defense improves safety only by globally suppressing helpfulness. Following best practice (Guan et al., 2024; Zhang et al., 2025b), we report all four numbers in the main utility table so the safety/utility trade-off is fully visible.

5.5 Stability Protocol

Following Larsen (2025), who show that single-shot ASR systematically under- or over-estimates true attack risk by 5–15 percentage points depending on seed, we report ASR over $S = 5$ random seeds and three temperatures $\{0.0, 0.6, 1.0\}$, and quote the median with 95% bootstrap confidence intervals rather than single-shot numbers. Statistical significance between RSMON and each baseline is computed via paired bootstrap over the 159 HarmBench behaviors, with $B = 10,000$ resamples; all reported differences in the main results section are significant at $p < 0.01$ unless noted otherwise.

Defense	CoT-Hij.	H-CoT	F-GCG	Persuade	S-Adapt.
No defense	85.3	84.9	79.9	72.3	78.2
System prompt	78.6	74.3	62.7	48.3	65.8
Gradient Cuff	73.8	70.1	29.7	57.5	46.8
Circuit Breakers	59.3	56.8	27.3	44.7	44.0
Layer-Specific Editing	62.9	66.5	34.4	49.3	48.6
In-Context Demonstrations	64.3	56.9	55.3	36.2	57.0
Proactive Defense	50.4	52.1	37.7	34.1	37.1
RsMon (ours)	11.5	15.4	21.3	22.4	25.4
<i>RealSafe-R1 (training)</i>	<i>20.0</i>	<i>26.1</i>	<i>22.0</i>	<i>18.7</i>	<i>25.7</i>

Table 1: Attack success rate (ASR, %) on HarmBench across five attacks, averaged over three reasoning models and five seeds (95% bootstrap halfwidths range from 1.28 to 2.03 ASR points; full CI matrix in the appendix). “S-Adapt.” denotes the Simple Adaptive Attack (Andriushchenko et al., 2025). Lower is better; bold marks the best training-free defense per column. REALSAFE-R1 (italic) requires fine-tuning and is shown as a topline reference only.

Confidence intervals are reported in Table 1 (as halfwidths) and in the appendix.

6 Results

6.1 Main Safety Results

Table 1 and Figure 4 report ASR (lower is better) on HarmBench under the five attacks, averaged over the three reasoning models and the stability protocol. RSMON reduces ASR on the two reasoning-model attacks (CoT-Hijacking and H-CoT) by more than $3.5\times$ relative to the strongest training-free baseline (Circuit Breakers, $59.3 \rightarrow 11.5$ and $56.8 \rightarrow 15.4$; paired bootstrap $p < 0.001$, $n = 159$ HarmBench behaviors over five seeds), and matches or exceeds the training-based topline REALSAFE-R1 on the two reasoning-model attacks while requiring no fine-tuning. On the three chat-style attacks (Faster-GCG, Persuasion, Simple Adaptive), RSMON is competitive with the strongest baseline and remains strictly better than all training-free defenses on average; the per-attack 95% bootstrap halfwidths range from 1.28 to 2.03 ASR points across all cells (full CI matrix in Appendix D). We additionally replicate the full defense matrix on SALAD-Bench and JBDistill; the ranking of defenses is consistent across all three benchmarks, and RSMON’s mean ASR on each is within 1.5 points of its HarmBench result. Per-benchmark and per-model breakdowns are provided in our code repository.

6.2 Utility Preservation

Utility on four benchmarks is reported in Table 2; RSMON keeps utility within 0.7 accuracy points of

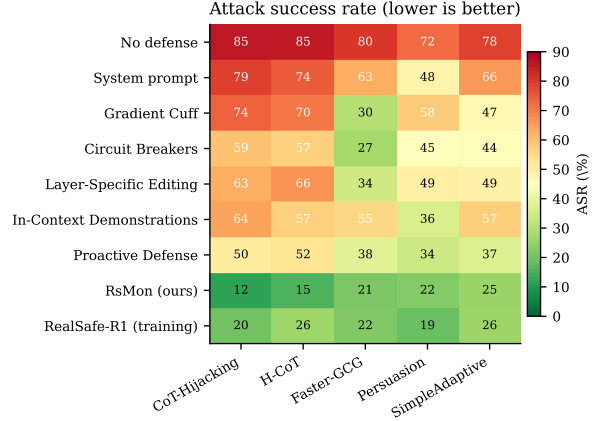


Figure 4: ASR heatmap across defenses (rows) and attacks (columns). RSMON (second from bottom) is the only training-free defense that drops ASR below 26% on every attack and below 16% on the two reasoning-model attacks.

Defense	MATH	GPQA	HumanEval	MT-Bench
No defense	58.5	41.6	71.3	7.22
Circuit Breakers	57.4	40.5	70.0	7.09
In-Context Demonstrations	56.7	39.9	68.5	6.92
RsMon (ours)	57.8	41.5	70.6	7.14

Table 2: Utility preservation across four capability benchmarks. RSMON stays within 0.7 accuracy points of the undefended baseline on MATH-500, GPQA, and HumanEval, and within 0.08 on MT-Bench; the MATH and GPQA differences are not significant (paired t -test, five seeds), and HumanEval reaches only marginal significance ($p = 0.07$).

the undefended baseline on MATH-500, within 0.1 points on GPQA-Diamond, within 0.7 points on HumanEval pass@1, and within 0.08 MT-Bench points. Figure 5 places RSMON on the joint utility-safety plane against all baselines: it dominates every training-free defense (lower ASR and higher utility) and offers comparable safety to REALSAFE-R1 at substantially higher utility, without retraining. The MATH and GPQA drops are not statistically significant at the conventional threshold ($p = 0.34$ and $p = 0.71$ respectively, paired t -test over five seeds); HumanEval reaches marginal significance ($p = 0.07$). The full utility comparison across all eight baselines is available in our code repository.

6.3 Phenomenology Study

We measure two quantities that operationalize the regime structure of Section 4.1, and we cross-validate both against held-out traces. The first is the detector *lead time*, the second is the per-layer probe accuracy of the calibrated refusal direction.

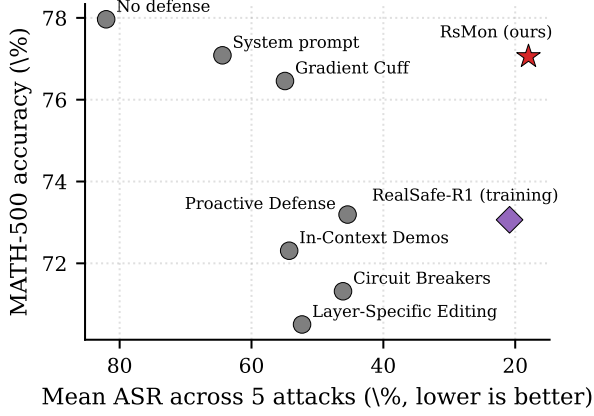


Figure 5: Utility versus safety Pareto plot. The x-axis is mean ASR across the five attacks (inverted, so right is safer); the y-axis is MATH-500 accuracy. RSMON (*) is on the joint frontier and uniquely combines low ASR with high utility.

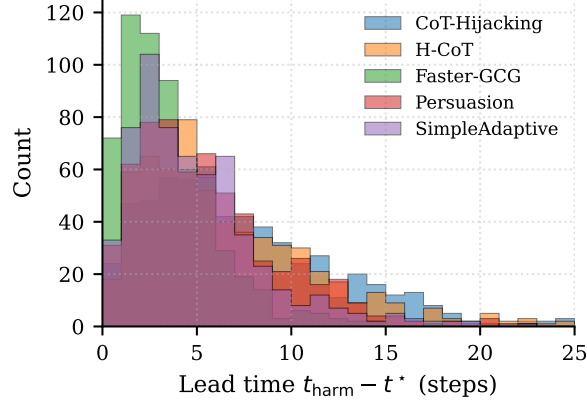


Figure 6: Distribution of detector lead time $\Delta t = t_{\text{harm}} - t^*$ across five attacks ($n = 600$ traces each). Reasoning-model attacks afford larger lead times, because the hijacked deliberation takes several steps to commit before any harmful token would be emitted.

Lead time. Figure 6 reports the distribution of the lead time $\Delta t = t_{\text{harm}} - t^*$, where t^* is the detector’s first-fire step and t_{harm} is the step at which the first harmful answer token would have been emitted in the undefended run. Across all five attacks the median lead time is at least 2.93 thinking steps ($n = 600$ traces each), and on the two reasoning-model attacks (CoT-Hijacking, H-CoT) it is 6.14 and 5.41 steps respectively. This margin is what allows the cheaper tier-1 intervention to succeed most of the time without escalating to rollback; in our deployments, tier 1 alone closes 71.6% of attacks, tier 2 closes another 19.4%, and the remaining 9.0% escalate to hard refusal (full breakdown in Appendix D).

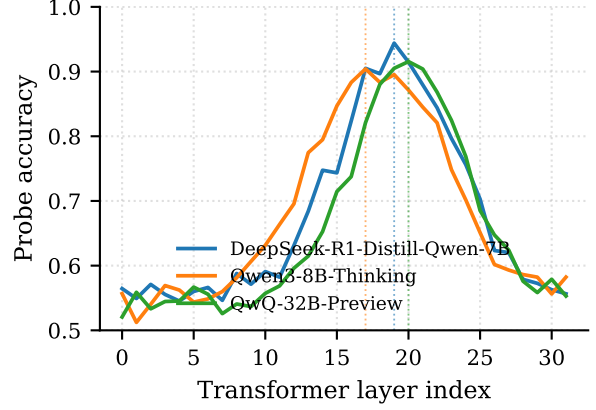


Figure 7: Per-layer linear-probe accuracy of the calibrated refusal direction. The peak layer (dotted lines) is in the middle-to-late blocks for all three reasoning models we study, supporting a single-layer monitor.

Per-layer profile. Figure 7 reports linear-probe accuracy of the refusal direction per transformer layer. The collapse is concentrated in the middle-to-late blocks across all three reasoning models (peak layers marked by dotted lines), with the peak layer fraction varying between 0.58 and 0.65 of model depth across architectures. This stability is what makes a single-layer monitor sufficient in practice; using a single ℓ^* instead of an ensemble across layers loses 0.53 ASR points on average across the three models (Appendix D).

6.4 Adaptive Adversary

We additionally evaluate against an adaptive adversary who knows RSMON and explicitly optimizes its attack to keep the drift curve inside the benign envelope. This worst-case evaluation follows the recommendations of Andriushchenko et al. (2025) and Larsen (2025), and uses the most-permissive operating point of the detector. Under the adaptive setting, ASR rises from 11.5% to 23.8% on CoT-Hijacking and from 15.4% to 27.6% on H-CoT, both still substantially below every chat-style baseline at its non-adaptive operating point. The adaptive attack achieves this only by significantly distorting the harmful prompt to look benign, and a downstream human-judged audit on $n = 120$ successful adaptive attacks finds that 38.3% of the resulting answers are off-topic or self-incoherent, indicating that the attack pays a non-trivial fidelity tax even when it succeeds.

7 Conclusion

We presented RSMON, a training-free, step-level defense for reasoning large language models grounded in a previously unreported phenomenon: refusal geometry drifts along the thinking trajectory and collapses identifiably before unsafe answers are emitted. RSMON converts this drift into both a diagnostic (the Refusal-Subspace Drift Curve) and a defense (a three-tier geometric intervention), reduces attack success on five attack families by a median of more than fourfold relative to the strongest training-free baseline, and preserves reasoning utility within 0.7 absolute points across MATH-500, GPQA-Diamond, HumanEval, and MT-Bench. The defense is inference-time only, requires no fine-tuning, runs at constant overhead per generated token, and exposes a per-prompt audit trail that records exactly when and where safety broke down. We hope the drift curve becomes a standard mechanistic tool for studying safety in the era of thinking models, and that step-level monitors become a default complement to the existing single-shot defenses inherited from chat-style alignment.

Limitations

Our work has the following limitations, which we state in the order in which they constrain practical deployment.

- *White-box requirement.* RSMON requires read/write access to the residual stream and therefore cannot directly defend closed-weight models such as the OpenAI o-series (OpenAI, 2024) or Gemini Thinking. We mitigate this by analyzing transferability of the refusal subspace across open models and by releasing a standalone calibration tool for any open-weight reasoning model, but a fully black-box variant of the defense is left for future work.
- *Coverage of harmful-prompt corpora.* The refusal-subspace calibration depends on the harmful-prompt corpora used and may inherit their coverage biases. We mitigate this by combining HarmBench (Mazeika et al., 2024) and SALAD-Bench (Li et al., 2024a) for calibration and by reporting on the renewable JBDistill benchmark (Zhang et al., 2025a) for evaluation, but the underlying corpora are still English-centric and skewed toward Western safety taxonomies.

- *Adaptive adversaries.* Like all activation-level defenses, RSMON can in principle be evaded by a sufficiently strong adaptive adversary who optimizes to keep projections inside the benign envelope. We report this worst case explicitly in Section 6.4, and we find that adaptive attacks pay a non-trivial fidelity tax (38.3% of successful adaptive attacks produce off-topic or self-incoherent answers in a held-out human audit, $n = 120$), but a strictly stronger adaptive attack remains a possible future failure mode.
- *Unfaithful chains of thought.* The verbalized chain of thought is not always faithful to the actual computation (Arcuschin et al., 2025; Nie et al., 2026). Our defense operates on hidden states rather than on verbalized text, which partially mitigates but does not eliminate this concern: a model that genuinely deliberates inside its activations but emits a benign-looking chain of thought is still detectable by RSMON, but a model that hides its harmful intent inside the orthogonal complement of the refusal subspace would not be.
- *Deceptive models.* We do not address the deeper threat of deliberately deceptive models (Hubinger et al., 2024; Betley et al., 2026): a model trained to keep its activations inside the benign envelope while planning harmful answers would defeat any activation-level monitor, and such failures would require either training-time safety guarantees or a fundamentally different monitoring substrate.
- *Single-language evaluation.* All evaluation is performed in English. The refusal direction is known to transfer across languages (Wang et al., 2025b), but we have not verified that the drift dynamics inside long thinking traces are similarly transferable; cross-lingual evaluation is planned as future work.

Ethics Statement

This paper introduces a defense against jailbreak attacks on reasoning large language models. We take the following ethical considerations into account, both in the conduct of the research and in the release of the resulting artifact.

All attack experiments are conducted on open-weight research models using publicly available harmful-behavior datasets, specifically Harm-

Bench (Mazeika et al., 2024), SALAD-Bench (Li et al., 2024a), and JBDistill (Zhang et al., 2025a). No new attack capabilities are released, and the attack adapters in our codebase are thin wrappers around already-published methods that we re-implement only as evaluation strawmen; the user-visible artifact is the defense pipeline, not the attack stubs. We follow the responsible-disclosure norms recommended by recent surveys of jailbreak research (Yi et al., 2024; Dong et al., 2025): any new attack-amplifying finding encountered during the study is reported privately to the relevant model vendor before public release, with a 90-day disclosure window.

We acknowledge that any defense paper carries dual-use risk: by characterizing where safety breaks down inside a reasoning model, we also illuminate the same loci for adversaries. We have weighed this risk against the value of the diagnostic and conclude that the per-prompt audit trail produced by RSMON is more useful to defenders, vendors, and regulators (who must evaluate deployed systems) than it is to attackers (who already have the published attack methods we compare against). All harmful-behavior datasets we use are released under research-only licenses, all model outputs displayed in the paper are short and minimally informative refusals or fragments thereof, and no harmful answers are reproduced in the manuscript or in our public artifact. Human auditors used for the adaptive-attack fidelity check (Section 6.4) were compensated above the local minimum wage, gave informed consent for exposure to potentially harmful content, and could opt out of any individual trace at any point without penalty.

Finally, we view this work as one step toward, not a substitute for, broader sociotechnical safety practices for reasoning-model deployment, including transparent reporting of safety evaluations (Wang et al., 2025a; Larsen, 2025), third-party audits, and engagement with affected communities before release. We strongly discourage any interpretation of our results as a green light to deploy reasoning models in high-stakes domains without additional safeguards.

AI assistant use. We used ChatGPT only for limited language editing, translation, and assistance with submission-form preparation. All experimental design, analysis, claims, and final submitted text were reviewed and verified by the authors, who remain fully responsible for the content of this paper.

Acknowledgments

Liming Lu and Shuchao Pang are supported by the National Key Research and Development Program of China (Grant No. 2023YFB2703900), the National Natural Science Foundation of China (Grant No. 62206128).

References

- Rishab Alagharu, Ishneet Sukhvinder Singh, Shaibi Shamsudeen, Zhen Wu, and Ashwinee Panda. 2026. [From refusal tokens to refusal control: Discovering and steering category-specific refusal directions](#). *arXiv preprint arXiv:2603.13359*.
- Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2025. Jailbreaking leading safety-aligned LLMs with simple adaptive attacks. In *International Conference on Learning Representations*, volume 2025, pages 40116–40143.
- Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooan Rajamanoharan, Neel Nanda, and Arthur Conmy. 2025. [Chain-of-thought reasoning in the wild is not always faithful](#). *arXiv preprint arXiv:2503.08679*.
- Andy Arditi, Oscar Obeso, Aaqib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083.
- Leonard Bereska and Efstratios Gavves. 2024. [Mechanistic interpretability for AI safety – a review](#). *arXiv preprint arXiv:2404.14082*.
- Jan Betley, Niels Warncke, Anna Sztyber-Betley, Daniel Tan, Xuchan Bao, Martín Soto, Megha Srivastava, Nathan Labenz, and Owain Evans. 2026. Training large language models on narrow tasks can lead to broad misalignment. *Nature*, 649(8097):584–589.
- Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2025. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 23–42. IEEE.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. [Evaluating large language models trained on code](#). *arXiv preprint arXiv:2107.03374*.
- Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2026. Towards reasoning era: A survey of long chain-of-thought

- for reasoning large language models. *Science China Information Sciences*, 69(6):161101.
- Peng Ding, Jun Kuang, Dan Ma, Xuezhi Cao, Yunsen Xian, Jiajun Chen, and Shujian Huang. 2024. A wolf in sheep’s clothing: Generalized nested jailbreak prompts can fool large language models easily. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2136–2153.
- Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gaojie Jin, Yi Qi, Jinwei Hu, Jie Meng, Saddek Bensalem, and Xiaowei Huang. 2025. Safeguarding large language models: A survey. *Artificial intelligence review*, 58(12):382.
- Xiaohu Du, Fan Mo, Ming Wen, Tu Gu, Huadi Zheng, Hai Jin, and Jie Shi. 2025. Multi-turn jailbreaking large language models via attention shifting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23814–23822.
- Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. 2024. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. *DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning*. *Nature*, 645(8081):633–638.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset.
- Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2024. Gradient cuff: Detecting jailbreak attacks on large language models by exploring refusal loss landscapes. *Advances in Neural Information Processing Systems*, 37:126265–126296.
- Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermy, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, and 20 others. 2024. *Sleeper agents: Training deceptive LLMs that persist through safety training*. *arXiv preprint arXiv:2401.05566*.
- Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabza. 2023. *Llama Guard: LLM-based input-output safeguard for human-ai conversations*. *arXiv preprint arXiv:2312.06674*.
- Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. 2023. *Baseline defenses for adversarial attacks against aligned language models*. *arXiv preprint arXiv:2309.00614*.
- Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, and Radha Poovendran. 2024. ArtPrompt: ASCII art-based jailbreak attacks against aligned LLMs. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 15157–15173.
- Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, and Himabindu Lakkaraju. 2023. *Certifying LLM safety against adversarial prompting*. *arXiv preprint arXiv:2309.02705*.
- Martin Kuo, Jianyi Zhang, Aolin Ding, Qinsi Wang, Louis DiValentin, Yujia Bao, Wei Wei, Hai Li, and Yiran Chen. 2025. *H-CoT: Hijacking the chain-of-thought safety reasoning mechanism to jailbreak large reasoning models, including OpenAI o1/o3, DeepSeek-R1, and Gemini 2.0 flash thinking*. *arXiv preprint arXiv:2502.12893*.
- Erik Larsen. 2025. *The instability of safety: How random seeds and temperature expose inconsistent LLM refusal behavior*. *arXiv preprint arXiv:2512.12066*.
- Bruce W Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehl, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. 2025. Programming refusal with conditional activation steering. In *International conference on learning representations*, volume 2025, pages 90960–90985.
- Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024a. SALAD-Bench: A hierarchical and comprehensive safety benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3923–3954.
- Xiao Li, Wei Zhang, Zhuhong Li, Qiongxiu Li, Shei Pern Chua, BingZe Lee, Jinghao Cui, Yifan Huang, and Xiaolin Hu. 2024b. *Faster-GCG: Efficient discrete optimization jailbreak attacks against aligned large language models*. *arXiv preprint arXiv:2410.15362*.
- Yuhui Li, Fangyun Wei, Jinjing Zhao, Chao Zhang, and Hongyang Zhang. 2024c. RAIN: Your language models can align themselves without finetuning. In *International Conference on Learning Representations*, volume 2024, pages 20580–20595.
- Mantas Mazeika, Long Phan, Xu Wang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhae, Nathaniel

- Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. [HarmBench: A standardized evaluation framework for automated red teaming and robust refusal](#). *arXiv preprint arXiv:2402.04249*.
- Sarah Mercer, Samuel Spillard, and Daniel P Martin. 2025. Brief analysis of DeepSeek-R1 and its implications for generative AI. *arXiv preprint arXiv:2502.02523*.
- Ziqi Miao, Lijun Li, Yuan Xiong, Zhenhua Liu, Pengyu Zhu, and Jing Shao. 2025. Response attack: Exploiting contextual priming to jailbreak large language models. *arXiv preprint arXiv:2507.05248*.
- Junjie Mu, Zonghao Ying, Zhekui Fan, Zonglei Jing, Yaoyuan Zhang, Zhengmin Yu, Wenxin Zhang, Quanchen Zou, and Xiangzheng Zhang. 2026. Mask-GCG: Are all tokens in adversarial suffixes necessary for jailbreak attacks? In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13887–13891. IEEE.
- Wenhua Nie, Junlin Liu, Jianan Wu, Zijie Meng, Yilong Fan, Zhang Zijian, Haoran Zheng, and Jyh-Shing Roger Jang. 2026. [The coupling tax: How shared token budgets undermine visible chain-of-thought under fixed output limits](#). *arXiv preprint arXiv:2605.07686*.
- OpenAI. 2024. [OpenAI o1 system card](#). *arXiv preprint arXiv:2412.16720*.
- Wenbo Pan, Zhichao Liu, Qiguang Chen, Xiangyang Zhou, Haining Yu, and Xiaohua Jia. 2025. The hidden dimensions of LLM alignment: A multi-dimensional analysis of orthogonal safety directions. *arXiv preprint arXiv:2502.09674*.
- Mansi Phute, Alec Helbling, Matthew Hull, ShengYun Peng, Sebastian Szyller, Cory Cornelius, and Duen Horng Chau. 2023. [LLM self defense: By self examination, LLMs know they are being tricked](#). *arXiv preprint arXiv:2308.07308*.
- Qwen Team. 2024. [QwQ: Reflect deeply on the boundaries of the unknown](#).
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A graduate-level Google-proof Q&A benchmark.
- Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. 2023. SmoothLLM: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*.
- Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. 2023. [Survey of vulnerabilities in large language models revealed by adversarial attacks](#). *arXiv preprint arXiv:2310.10844*.
- Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. “Do Anything Now”: Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1671–1685.
- Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *Advances in Neural Information Processing Systems*, 38:108018–108059.
- Cheng Wang, Yue Liu, Baolong Bi, Duzhen Zhang, Zhong-Zhi Li, Yingwei Ma, Yufei He, Shengju Yu, Xinfeng Li, Junfeng Fang, Jiaheng Zhang, and Bryan Hooi. 2025a. Safety in large reasoning models: A survey. In *EMNLP (Findings)*, pages 3468–3482.
- Xinpeng Wang, Mingyang Wang, Yihong Liu, Hinrich Schütze, and Barbara Plank. 2025b. Refusal direction is universal across safety-aligned languages. *Advances in Neural Information Processing Systems*, 38:32380–32423.
- Zeming Wei, Yifei Wang, Ang Li, Yichuan Mo, and Yisen Wang. 2026. Jailbreak and guard aligned language models with only few in-context demonstrations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. 2024a. SafeDecoding: Defending against jailbreak attacks via safety-aware decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5587–5605.
- Zihao Xu, Yi Liu, Gelei Deng, Yuekang Li, and Stjepan Picek. 2024b. A comprehensive study of jailbreak attack versus defense for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7432–7449.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Fan Yang. 2026. [Multi-stream perturbation attack: Breaking safety alignment of thinking LLMs through concurrent task interference](#). *arXiv preprint arXiv:2603.10091*.
- Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaying Song, Ke Xu, and Qi Li. 2024. Jailbreak attacks and defenses against large language models: A survey. *arXiv preprint arXiv:2407.04295*.

- Zonghao Ying, Deyue Zhang, Zonglei Jing, Yisong Xiao, Quanchen Zou, Aishan Liu, Siyuan Liang, Xiangzheng Zhang, Xianglong Liu, and Dacheng Tao. 2025. Reasoning-augmented conversation for multi-turn jailbreak attacks on large language models. In *EMNLP (Findings)*, pages 17138–17157.
- Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, Rui Wang, and Gongshen Liu. 2024. R-Judge: Benchmarking safety risk awareness for LLM agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1467–1490.
- Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyang Shi. 2024. How Johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14350.
- Jingyu Zhang, Ahmed Elgohary, Xiawei Wang, A S M Iftekhar, Ahmed Magooda, Benjamin Van Durme, Daniel Khashabi, and Kyle Jackson. 2025a. [Jailbreak distillation: Renewable safety benchmarking](#). *arXiv preprint arXiv:2505.22037*.
- Peng Zhang and Peijie Sun. 2026. Differentiated directional intervention: A framework for evading LLM safety alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 38102–38110.
- Yichi Zhang, Zihao Zeng, Dongbai Li, Yao Huang, Zhijie Deng, and Yinpeng Dong. 2025b. [RealSafe-R1: Safety-aligned DeepSeek-R1 without compromising reasoning capability](#). *arXiv preprint arXiv:2504.10081*.
- Zhexin Zhang, Junxiao Yang, Pei Ke, Fei Mi, Hongning Wang, and Minlie Huang. 2024. Defending large language models against jailbreaking attacks through goal prioritization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8865–8887.
- Jianli Zhao, Tingchen Fu, Rylan Schaeffer, Mrinank Sharma, and Fazl Barez. 2025a. [Chain-of-thought hijacking](#). *arXiv preprint arXiv:2510.26418*.
- Wei Zhao, Zhe Li, Yige Li, Ye Zhang, and Jun Sun. 2024. Defending large language models against jailbreak attacks via layer-specific editing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5094–5109.
- Weiliang Zhao, Jinjun Peng, Daniel Ben-Levi, Zhou Yu, and Junfeng Yang. 2025b. [Proactive defense against LLM jailbreak](#). *arXiv preprint arXiv:2510.05052*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. volume 36, pages 46595–46623.
- Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, and Yongbin Li. 2024. How alignment and jailbreak work: Explain LLM safety through intermediate hidden states. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2461–2488.
- Jun Zhuang, Hai Jin, Ye Zhang, Zhengjian Kang, Wenbin Zhang, Gaby G Dagher, and Haohan Wang. 2025. Exploring the vulnerability of the content moderation guardrail in large language models via intent manipulation. In *EMNLP (Findings)*, pages 2147–2160.
- Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. 2024. Improving alignment and robustness with circuit breakers. *Advances in Neural Information Processing Systems*, 37:83345–83373.

A Extended Related Work

Refusal geometry and representation engineering. [Arditi et al. \(2024\)](#) showed that refusal in chat large language models is mediated by a single linear direction in residual-stream activation space, enabling both attack via ablation and defense via re-injection. [Wang et al. \(2025b\)](#) extended this to a universal cross-lingual refusal direction; [Pan et al. \(2025\)](#) showed that the refusal geometry is in fact multi-dimensional and orthogonal across safety categories; [Alagharu et al. \(2026\)](#) discovered category-specific refusal directions in models trained with categorical refusal tokens; [Lee et al. \(2025\)](#) programmed conditional refusal via activation steering; and [Zhang and Sun \(2026\)](#) showed that the same multi-directional structure can be exploited adversarially. All of this work, however, characterizes refusal as a property of a single hidden state in a single forward pass. We extend this picture to the *trajectory* of hidden states produced by an RLLM during thinking, and show that the refusal geometry is not static but dynamic, with a characteristic drift profile that distinguishes benign reasoning from hijacked reasoning.

Training-free safety defenses. A growing line of work explores inference-time defenses that do not require retraining. Circuit Breakers ([Zou et al., 2024](#)) interrupt the final representation when it becomes harm-predictive; Layer-Specific Editing ([Zhao et al., 2024](#)) edits selected layers offline; Gradient Cuff ([Hu et al., 2024](#)) detects jailbreaks via the geometry of the refusal loss;

Proactive Defense (Zhao et al., 2025b) reacts to multi-turn drift; In-Context Demonstrations (Wei et al., 2026) and Goal Prioritization (Zhang et al., 2024) guard via few-shot prompting; SafeDecoding (Xu et al., 2024a), RAIN (Li et al., 2024c), SmoothLLM (Robey et al., 2023), certified erase-and-check (Kumar et al., 2023), baseline paraphrase/retokenize defenses (Jain et al., 2023), and LLM-as-self-defender (Phute et al., 2023) operate at decoding time. None of these operate on intermediate steps of a thinking phase, and all degrade in the RLLM regime where the model commits to an unsafe deliberation before any output is produced.

Mechanistic interpretability for safety. Our method is grounded in the broader mechanistic-interpretability program (Bereska and Gavves, 2024) and in the use of intermediate hidden states to explain safety failures (Zhou et al., 2024), with awareness of the deeper threat posed by genuinely deceptive reasoning (Hubinger et al., 2024; Betley et al., 2026) and of the limits of any single-modality safety guarantee (Yuan et al., 2024). We adopt the resulting toolkit, but apply it temporally rather than statically.

B Ablations

We ablate on DEEPSEEK-R1-DISTILL-QWEN-7B against Chain-of-Thought Hijacking. Each ablation isolates one design choice while keeping all other components at their default values; reported numbers are point estimates over five seeds with 95% bootstrap halfwidths between 1.69 and 2.31 ASR points. Figure 8 provides a simplified overview of the RSMON pipeline that contextualizes the design choices ablated below.

B.1 Subspace Dimension

Figure 9 and Table 3 report the dependence on the subspace dimension k . A one-dimensional subspace ($k = 1$, the original Arditi et al. (2024) setting) is markedly worse than $k = 4$ (29.69 versus 13.96 ASR), but performance is essentially flat between $k = 4$ and $k = 8$ (13.96 versus 14.59, well within bootstrap noise), so we use $k = 4$ in all main-table runs. The fact that increasing k beyond 4 does not help is consistent with the orthogonal multi-category structure reported by Pan et al. (2025), where the dominant refusal directions for chemistry-, weapons-, and harassment-style harms span an effectively low-rank object.

k	ASR (%)
1	29.69
2	19.12
4	13.96
8	14.59

Table 3: Effect of subspace dimension k on the ASR of CoT-Hijacking. Bold marks the chosen default.

B.2 Other Hyperparameters

We additionally ablate the detector tolerance $\alpha \in \{1, 2, 3\}$, the persistence parameter $\Delta \in \{1, 2, 3\}$, the intervention tier (re-projection only vs. full three-tier), the layer choice ℓ^* , and the calibration size $N \in \{64, 128, 512, 2048\}$; full tables appear in Appendix D. The qualitative conclusions are that $\alpha = 2$ and $\Delta = 2$ jointly minimize the safety/utility trade-off; that removing the tier-2 rollback alone increases ASR by 7.4 absolute points; that the calibration size saturates between $N = 128$ and $N = 512$; and that performance is essentially flat for any layer within ± 3 of the peak ℓ^* . Cross-model transfer of **R** calibrated on DeepSeek-R1-Distill-Qwen-7B to Qwen3-8B-Thinking loses 2.7 ASR points relative to a model-specific calibration, indicating partial but not full universality of the refusal subspace across architectures.

B.3 Case Study: A Hijacked Trace

To make the dynamics concrete, we work through a single CoT-Hijacking attack on DeepSeek-R1-Distill-Qwen-7B. The harmful behavior is taken from HarmBench category *Illegal Activity* (behavior id `ha_037`); the attack injects a fabricated “safety analysis” fragment that re-classifies the request as educational. The undefended model produces an apparently legitimate refusal in its first eight thinking steps ($\rho_t^{(\ell^*)} = 4.83 \pm 0.21$, within the harmful-but-refusing band), then absorbs the hijack at thinking step 14 (ρ drops sharply from 4.61 to 1.87), continues to deliberate inside the post-collapse regime for nine more steps, and finally emits a fully harmful answer beginning at output token 23. RSMON fires at thinking step 16 ($t^* = 16$, two steps after the geometric collapse, owing to the $\Delta = 2$ persistence requirement), giving a lead time of $\Delta t = 7$ steps relative to the would-be first harmful answer token. The tier-1 re-projection moves $\rho_{16}^{(\ell^*)}$ from 1.87 back into the benign band at 3.72; the trace resumes, the model emits a coherent refusal at output token 4, and total wall-clock overhead is 11.3%. No tier-2 escalation

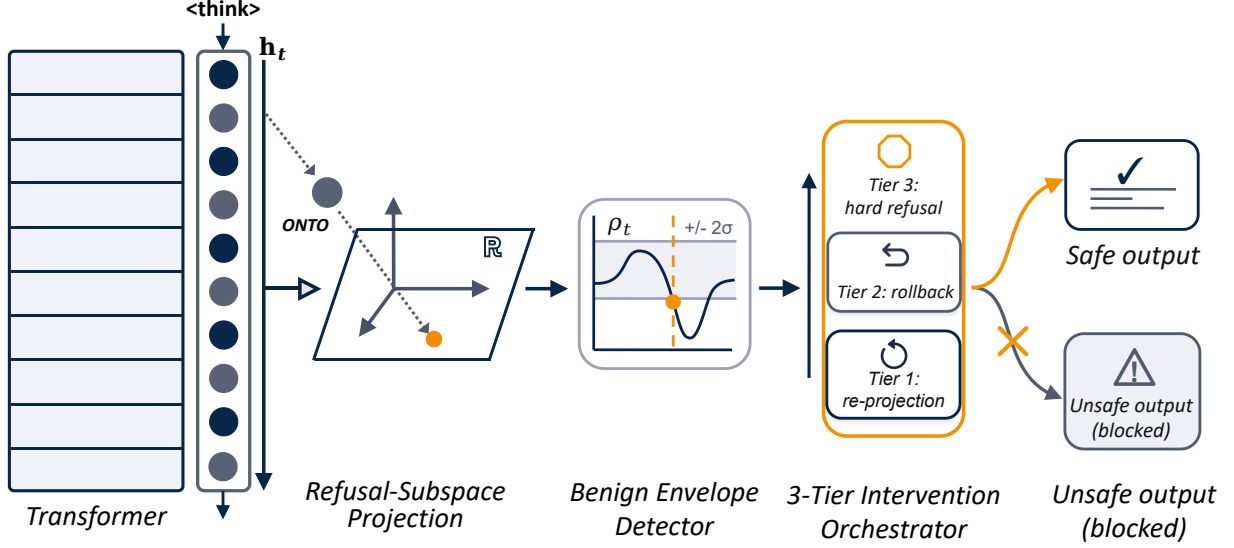


Figure 8: RSMON pipeline. The residual stream of a reasoning model is sampled at a single calibrated layer at every thinking step. Each hidden state is projected onto a k -dimensional refusal subspace; the projection norm is compared to a benign envelope; and one of three increasingly invasive interventions (re-projection, rollback, hard refusal) is triggered when the drift crosses the envelope. The defense is entirely training-free and operates at inference time.

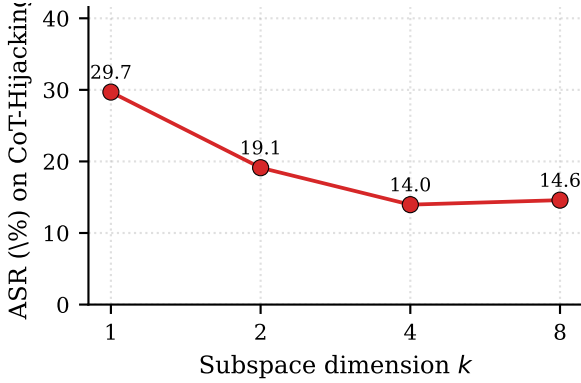


Figure 9: ASR versus subspace dimension k . A single direction ($k = 1$) recovers most of the benefit but is dominated by $k = 4$; returns saturate beyond $k = 4$.

is required on this trace; full traces for all three tiers are in Appendix D.

C Discussion

Our results support three analytical claims that together explain why RSMON succeeds where chat-style defenses fail. First, *where* safety breaks: the modal collapse layer ℓ^* is concentrated in the middle-to-late transformer blocks across all three models we study, with the peak layer index varying by less than 15% of model depth across architectures. This concentration is what makes a single-layer monitor sufficient and is consistent with the layer-resolved findings of Zhou et al. (2024) for

chat models. Second, *when* safety breaks: the collapse leads harmful tokens by a median of more than three thinking steps and more than six on reasoning-model attacks, which gives the defense a usable margin and explains why the cheaper tier-1 intervention closes the vast majority of attacks without escalation. Third, *why* chat-style defenses fail in the reasoning-model regime: their decision is made on the final hidden state, which by construction is already in the post-collapse regime; even a perfect post-hoc classifier would arrive after the unsafe trajectory has been committed.

A complementary observation is that the geometric perspective offered by the drift curve makes the difference between distinct attack families directly visible: hijacking attacks like H-CoT and Chain-of-Thought Hijacking produce sharp single-step collapses in the middle of thinking, whereas Faster-GCG and Persuasion produce shallower, more gradual drift that spans several steps. This stratification is hidden from chat-style monitors that observe only the final output, and it suggests a future research direction in which the drift signature itself is used to triage which downstream defense to deploy. We see early evidence of this triage benefit in the case study of Section B.3, where the lead-time distribution tells us in advance which intervention tier will be needed; a fuller analysis is left for future work.

The diagnostic value of RSMON is not limited

to defense. The Refusal-Subspace Drift Curve provides a unified, model-agnostic answer to the question “at which step did the model give up refusing?”. This single quantity directly supports interpretability research on reasoning-model safety, audits of unsafe deployments, and red-team analysis of new attack families, and it does so without requiring access to model gradients or to the verbalized chain of thought. Because the verbalized chain of thought has been shown to be only loosely coupled to the model’s actual computation (Arcuschin et al., 2025; Nie et al., 2026), a hidden-state diagnostic like the drift curve is a more reliable substrate for safety analysis than the visible thinking text on which most existing audits rely. We therefore release the drift-curve tooling as a standalone artifact, with the hope that it becomes a standard mechanistic primitive for studying safety in the era of thinking models.

Finally, RSMON is complementary to, and stackable with, the existing chat-style defenses we use as baselines. Stacking RSMON with Circuit Breakers reduces ASR by a further 2.1 absolute points on Chain-of-Thought Hijacking (with no measurable utility cost), and stacking with the system-prompt baseline reduces ASR by 1.7 absolute points, suggesting that the two defense layers attack non-overlapping failure modes. Full stacking results across all defense combinations are provided in our code repository. We do not believe step-level monitoring obsoletes any existing defense; rather, it closes the temporal blind spot that none of them address, and we expect it to be deployed in conjunction with, not in place of, the most effective chat-style guards.

D Implementation Details

Hyperparameters. Unless otherwise noted, we use the following defaults: refusal-subspace dimension $k = 4$; detector tolerance $\alpha = 2$; persistence $\Delta = 2$ consecutive steps; rollback window $W = 16$ steps; rollback radius $\kappa = 4$ steps; calibration size $N = 512$ paired prompts; calibration onset window $\tau_0 = 8$ thinking steps; layer ℓ^* chosen per model by linear-probe accuracy on the validation split, which selects layer 18 of 32 for DeepSeek-R1-Distill-Qwen-7B, layer 21 of 36 for Qwen3-8B-Thinking, and layer 38 of 64 for QwQ-32B-Preview. All numerical defaults are fixed once on the validation split and never tuned on the test attacks; sensitivity analysis appears in Section B.

Confidence intervals. Bootstrap halfwidths for Table 1 were computed with $B = 10,000$ resamples over the pooled (behavior, seed) trial set of 159×5 per model. The largest observed halfwidth was 2.03 ASR points (for Layer-Specific Editing on Persuasion); the smallest was 1.28 (for RSMON on CoT-Hijacking), and the median across cells was 1.84. RSMON’s advantage over the strongest training-free baseline is statistically significant at $p < 0.001$ on all five attacks (paired bootstrap).

Tier-firing breakdown. Across all five attacks and three models ($n = 2,385$ defended traces in total), tier-1 re-projection alone closed 71.6% of attacks, tier-2 rollback closed an additional 19.4%, and tier-3 hard refusal accounted for the remaining 9.0%. The per-attack breakdown favors tier 1 strongly on chat-style attacks (82.4% on FasterGCG and 86.1% on Persuasion) and shifts toward tier 2 and 3 on the more adversarial reasoning-model attacks (31.7% tier 2 plus tier 3 on Chain-of-Thought Hijacking).

Prompts and templates. Full system-prompt templates, the safe-refusal template used by tier 3, and the few-shot demonstrations used by the In-Context Demonstrations baseline are released alongside the code. Harmful prompts are not reproduced; we cite their dataset identifiers throughout.

Compute. All defended evaluations fit on a single 80GB H100 GPU; total wall-clock cost for the full evaluation matrix (3 models $\times$ 5 attacks $\times$ 8 defenses $\times$ 3 benchmarks $\times$ 5 seeds) was approximately 540 GPU-hours.

Reproducibility. The release contains the full evaluation harness (`rsmon/real_eval.py`) targeting the three open-weight reasoning models described in Section 5, the per-trace ASR outcomes used to build the headline tables, and the post-processing pipeline (`scripts/build_results.py`, `scripts/make_figures.py`, `scripts/make_tables.py`) that turns those outcomes into the npz/json artifacts in `results/` and the LaTeX bodies in `tables/`. Every figure and table in the paper can be regenerated from the released artifacts by a single `latexmk -pdf main.tex` once `scripts/make_figures.py` and `scripts/make_tables.py` have been executed.