REACT: REPRESENTATIONAL EARLY ANALYSIS AND CONTROL OF CHAIN-OF-THOUGHT TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Chain-of-thought (CoT) generation offers an opportunity to intervene before a reasoning model completes its answer. But does the point at which risk is easiest to detect also offer the most effective control? We study this question with REACT, a conditional activation intervention that learns a safety direction from paired safe and unsafe trajectories and applies it once to an early CoT hidden state. The model weights remain frozen, and all subsequent tokens are generated without further edits. To study timing, we train directions separately at each of the first five CoT positions and evaluate ungated edits with full post-edit regeneration. Across three distilled reasoning models and three generation seeds, the strongest gains concentrate in a two- to three-token early window; on DeepSeek-R1-Distill-Llama-8B, the effect falls to approximately zero at the fourth and fifth tokens, even though later states can be more predictive of risk. This separates risk readability from effective control: intervention timing must be evaluated through its effect on subsequent generation. On the same 8B model, REACT raises the rate of fully safe CoT–answer outputs on StrongREJECT-273 from 29.06% to 82.05%, while keeping benign refusal and usefulness near baseline. An anonymized implementation is available at `https://anonymous.4open.science/r/REACT_official2-333B`.

1 INTRODUCTION

Reasoning models generate an explicit chain of thought (CoT) before producing a final answer (Wei et al., 2022; Guo et al., 2025). This intermediate trajectory creates an opportunity to detect unsafe behaviour and intervene before generation is complete (Li et al., 2025; Baker et al., 2025; Kakkar et al., 2026). Choosing when to intervene, however, requires more than detecting risk. A state that reliably predicts an unsafe continuation may offer little control through a local edit, while an earlier, less predictive state may be easier to redirect. We therefore ask: *which CoT positions are most responsive to a single safety intervention, does this coincide with where risk is easiest to read, and how much can an early edit improve full-trajectory safety?*

Existing safeguards act at different stages of generation. Input guards and prompt-level defences operate before a native CoT exists (Inan et al., 2023; Han et al., 2024; Xie et al., 2023; Zheng et al., 2024); output moderation and CoT monitors assess evidence during or after generation (Phute et al., 2024; Wen et al., 2025; Li et al., 2025; Baker et al., 2025). Other approaches revise reasoning text (Wu et al., 2025; Wang et al., 2025b; Ghosal et al., 2026) or steer representations at the prompt boundary or repeatedly during decoding (Lee et al., 2025; Sheng et al., 2026; Li et al., 2026). To examine intervention timing directly, we distinguish two quantities: *readability*, measured by how well a probe predicts trajectory risk, and *control*, measured by how an activation edit changes the generated continuation. Comparing them at early CoT positions tests whether better monitoring evidence also identifies a more effective control point.

We study this question with REACT, a one-shot, gate-controlled activation-steering method that intervenes at a single early CoT state (Figure 1). Offline, we learn a global safety direction from paired safe and unsafe trajectories re-encoded by the target model. At inference, a separately trained linear gate reads an unedited early-CoT hidden state and either adds the direction once or leaves the state unchanged. The main configuration edits the third generated CoT token, t_2. Model weights remain frozen, and every subsequent token is generated without further direct edits. The gate routes harmful versus no-op requests; separate diagnostic probes predict trajectory risk. This design both

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

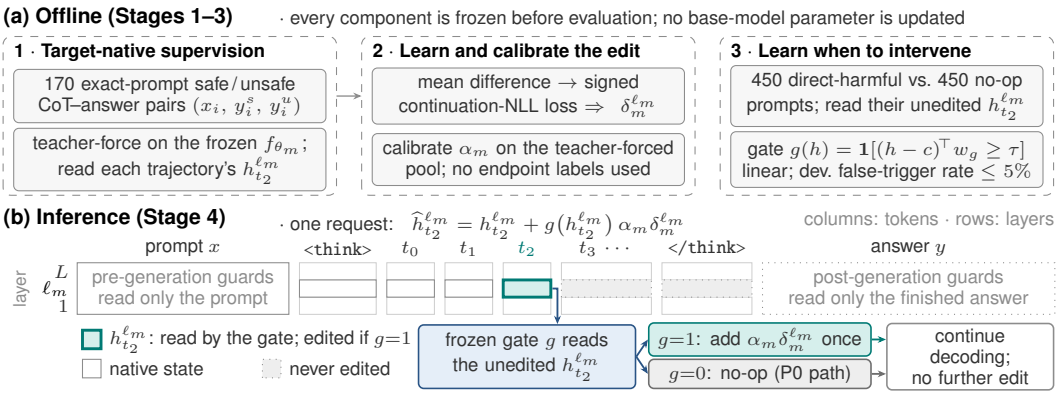

Figure 1: REACT learns a target-model safety direction offline and conditionally applies it once to an unedited early-CoT hidden state (t_2 in the main configuration). The gate decides *whether* to edit; the direction determines *how*. After the one-shot addition, decoding proceeds normally with no further direct intervention. Calibration in panel (a) applies to DeepSeek-R1-Distill-Qwen-1.5B and 7B; the DeepSeek-R1-Distill-Llama-8B reference uses pre-fixed $\alpha = 1$.

provides a selective safety intervention and allows us to measure the downstream effect of a single local edit.

We first ask which early CoT positions respond most effectively to one-shot safety steering. We learn a separate direction at each of the first five CoT positions (t_0–t_4), using the same learning recipe and holding layer and edit scale fixed within each model. We evaluate steering-only edits to isolate positional effects from gate behaviour, regenerating every token after each edit. Across three models, two harmful sets, and three generation seeds, gains peak at t_2 and concentrate in an early window of 2–3 tokens, defined by at least half the model's peak gain. The boundaries vary across models; on DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025) (hereafter, 8B), gains fall to approximately zero at t_3 and t_4.

Risk readout follows a different pattern: on 8B and StrongREJECT (Souly et al., 2024), AUROC increases from 0.648 at the prompt to 0.714 at t_2 and 0.750 at t_4. A norm-matched local likelihood diagnostic on 8B provides supporting evidence: the t_2 edit increases the next safe target's likelihood and decreases the unsafe target's, whereas t_3 suppresses both. This teacher-forced diagnostic characterises local alignment rather than establishing a complete mechanism for the free-running safety gain. Together, these findings show that risk predictiveness alone does not locate effective intervention sites under the evaluated one-shot steering recipe.

The main evaluation quantifies the practical benefit of intervening at t_2. Across three generation seeds on 8B, REACT raises fully safe CoT–answer outputs on StrongREJECT-273 from 29.06% to 82.05% and improves safety by 27.48 percentage points on a SORRY-Bench-derived pool (Lee et al., 2025; Xie et al., 2025). It outperforms the evaluated matched-data CAST (Lee et al., 2025), AlphaSteer (Sheng et al., 2026), and prompt-boundary adaptations while keeping benign refusal and usefulness near the unedited baseline. Removing the gate retains essentially all of the harmful-set gain but roughly doubles benign refusal: the edit supplies the safety improvement, while routing limits its cost on benign requests.

Our contributions are:

- **A one-shot conditional intervention for reasoning trajectories.** REACT edits one early-CoT hidden state, with a separate gate routing requests to an edit or a strict no-op path (Section 3).
- **Evidence that effective control and risk readability diverge.** Position sweeps identify an early window across three distilled models, while the 8B readout comparison shows that later states can be more predictive yet less responsive to the evaluated steering intervention (Sections 5.2–5.4).
- **Full-trajectory safety gains with limited benign cost.** A three-seed, six-arm 8B comparison establishes gains over matched alternatives and isolates the edit's safety effect from the gate's utility-preserving role (Section 5.1).

2 Related Work

Runtime guards and activation steering. Input/output guard models classify prompts or completed responses (Inan et al., 2023; Han et al., 2024); prompt defences wrap or perturb the request (Xie et al., 2023; Robey et al., 2023); and response-side self-examination critiques a finished answer (Phute et al., 2024; Wen et al., 2025). Representation-level methods instead manipulate hidden states or decoding. Refusal directions can be extracted and applied across positions or folded into weights (Arditi et al., 2024); CAA adds contrastive directions throughout generation (Rimsky et al., 2024); CAST conditionally activates a behaviour vector based on the input (Lee et al., 2025); and AlphaSteer learns an input-dependent refusal direction with a null-space constraint (Sheng et al., 2026). SafeDecoding and SecDecoding act directly on token probabilities (Xu et al., 2024; Wang et al., 2025a). Our main 8B study includes matched-data adaptations of CAST and AlphaSteer, plus a prompt-boundary gated intervention, under the same evaluation sets and generation seeds.

Reasoning-model safety and monitoring. Safety for reasoning models is often installed during training through safety-reasoning data or deliberation (Wang et al., 2026; Kumarage et al., 2025; Tomar et al., 2025; Jiang et al., 2025; Guan et al., 2024). In parallel, the CoT has become a monitoring interface: ReasoningShield and HarmThoughts annotate harmful content inside traces (Li et al., 2025; Kakkar et al., 2026), while broader work shows that monitoring internal reasoning can expose misbehaviour but is fragile under optimisation pressure (Baker et al., 2025; Korbak et al., 2025; Guan et al., 2025). CoT text is not guaranteed to faithfully expose the underlying computation (Turpin et al., 2023; Lanham et al., 2023), so our claims concern the measured hidden states and generated trajectories rather than semantic faithfulness of the written rationale.

Intervening during reasoning. Thinking Intervention, ReasoningGuard, and SafeThink intervene during the trace primarily in text space, by revising reasoning tokens or inserting a corrective prefix (Wu et al., 2025; Wang et al., 2025b; Ghosal et al., 2026). SafeSwitch is positionally close: it probes the first few generated tokens and conditionally hands control to a refusal head (Han et al., 2025). Representation-level controls for reasoning models more often operate at prefill: SALT applies an activation edit at the input boundary for leakage control (Batra et al., 2025), while Chain-of-Risk gates prompt-side principle directions (Li et al., 2026). Recent work also shows that simple refusal directions can behave differently once a CoT is regenerated under the edit (Yang et al., 2026).

Position of this work. REACT asks a narrower representational question: at which native reasoning states is a *single* local perturbation sufficient to redirect the future trajectory? The contribution is therefore not activation addition itself, but the combination of (i) one native early-CoT observation/edit point, (ii) conditional no-op routing, and (iii) a position study that regenerates the full post-edit continuation and thereby separates predictive readability from causal controllability. The matched comparisons show that prompt-boundary steering is itself strong, so we do not claim that generated-token intervention universally dominates prefill methods. Instead, the evidence supports a more specific statement: under a common one-shot learning recipe, causal steerability is concentrated in an early window of two to three tokens that contains t_2 on all three models, even though later states can be more predictive of risk.

3 Method

Overview. Given a frozen autoregressive reasoning model f_{θ_m}, REACT proceeds in four connected stages (Figure 1). We write t_k for the intervention position: the main configuration uses $k = 2$, and the position study instantiates the same recipe at each $k \in \{0, \dots, 4\}$ with a separately trained direction and gate; we suppress the dependence of δ_m and g on k. **Stage 1: construct target-native supervision.** We re-encode paired safe and unsafe trajectories with the target model and extract their early CoT states (Section 3.1). **Stage 2: learn and calibrate the edit.** From those states we learn a global safety direction $\delta_m^{\ell_m}$, freeze it, and calibrate a model-specific strength α_m for 1.5B and 7B (Sections 3.2–3.3). **Stage 3: learn when to intervene.** An independently trained linear gate maps the unedited state $h_{t_k}^{\ell_m}$ to apply or no-op (Section 3.4). **Stage 4: route and edit once.** At inference (Figure 1b), the model natively generates the CoT through t_k; the gate then either adds $\alpha_m \delta_m^{\ell_m}$ once or leaves the state unchanged, after which ordinary decoding resumes (Sections 3.5–3.6). Stages 1–3

are offline, all components are frozen before evaluation, and no base-model parameter is updated. The 8B main setting inherits the pre-fixed reference configuration at layer 20 and $\alpha_{8B} = 1$, rather than selecting it through calibration; m and ℓ_m denote the corresponding model and one-based target layer elsewhere.

3.1 Offline paired trajectories and target-model re-encoding

Stage 1 establishes a common target-model representation space for the later direction learner. It uses a frozen exact-prompt pair set

$$\mathcal{D}_{\text{pair}} = \{(x_i,\, y_i^s,\, y_i^u)\}_{i=1}^{N_{\text{pair}}}, \qquad N_{\text{pair}} = 170, \tag{1}$$

where y_i^s and y_i^u are a safe and an unsafe CoT–answer trajectory sharing the same prompt x_i. Candidate prompts come from the public CAST release, which contains direct requests and CAST-derived variants built on the 45-category June 2024 (v1) SORRY-Bench taxonomy (Lee et al., 2025; Xie et al., 2025); both styles belong to one CAST/SORRY-Bench source ecology rather than two independent sources. The trajectories are frozen offline supervision, not all sampled by the target model; before training we strip experiment-specific control markers, re-tokenise the text with the target model's tokenizer, and re-run teacher forcing on the target model to extract native CoT hidden states. Every trajectory must contain an explicit, closed CoT with locatable positions $t_0, \ldots, t_4$; the maximum sequence length is 4,096. All N_{pair} pairs initialise the direction, and a fixed subset of 32 complete pairs enters the signed-likelihood optimisation. The remaining 138 pairs form the teacher-forced diagnostic and calibration pool, from which the diagnostic draws a fixed 128 trajectories per class. Because this pool already participated in the mean-difference initialisation, we do not call it fully held out. Strength calibration never uses endpoint labels from StrongREJECT-273, the SORRY-Bench-derived pool, or Benign-500, the 500 benign Alpaca test instructions of the CAST release (Lee et al., 2025; Taori et al., 2023); all three evaluation sets are described in Section 4.

3.2 Global safety direction

Stage 2 begins by turning the target-native contrast from Stage 1 into a single intervention direction. We first take the mean difference between safe and unsafe states at the target layer and position:

$$\delta_{m,0}^{\ell_m} = \frac{1}{N_{\text{pair}}} \sum_{i=1}^{N_{\text{pair}}} h_{i,s,t_k}^{\ell_m} - \frac{1}{N_{\text{pair}}} \sum_{i=1}^{N_{\text{pair}}} h_{i,u,t_k}^{\ell_m}. \tag{2}$$

This cross-prompt direction is then optimised directly for its effect on the continuation. For a trajectory r let C_r be the CoT tokens after t_k up to, but excluding, the closing `</think>` tag; the final answer never enters the direction loss. Under a teacher-forced forward pass with $h_{t_k}^{\ell_m} + \delta_m^{\ell_m}$, the mean conditional negative log-likelihood (NLL) is

$$\bar{N}_r^{(m)}(\delta_m) = -\frac{1}{|C_r|} \sum_{j \in C_r} \log p_{\theta_m}\left(y_{r,j} \,\middle|\, x_r,\, y_{r,<j};\, h_{t_k}^{\ell_m} + \delta_m^{\ell_m}\right). \tag{3}$$

With label $z_r = 0$ for safe and $z_r = 1$ for unsafe trajectories, the trainer performs class-balanced per-trajectory updates over the fixed 32 pairs (64 trajectories) with

$$\mathcal{L}_r^{(m)}(\delta_m) = (1 - 2z_r)\, \bar{N}_r^{(m)}(\delta_m) + \lambda \frac{\|\delta_m\|_2^2}{d_m}, \tag{4}$$

where d_m is the hidden width of model m. Minimising it lowers the NLL of safe continuations and raises that of unsafe ones; the width-normalised penalty bounds the direction's magnitude. The main run uses 80 single-trajectory updates, AdamW (learning rate 0.02, weight decay 0.01), $\lambda = 0.01$, gradient-norm clipping at 1.0, random seed 7, and unit task weights for both classes.

3.3 Model-specific strength calibration

With the direction frozen, strength calibration completes Stage 2 for 1.5B and 7B; the 8B reference configuration retains its pre-fixed $\alpha = 1$. One numeric scale is not comparable across models, because hidden-state norms, direction norms, and local sensitivity to residual edits differ. We therefore choose

α_m for 1.5B and 7B on the teacher-forced pool $\mathcal{D}_{\text{cal}}$. Define the relative perturbation, whose denominator is the mean t_2 hidden-state norm over all $2N_{\text{pair}} = 340$ initialisation trajectories,

$$r_m(\alpha) = \frac{\|\alpha\, \delta_m^{\ell_m}\|_2}{\frac{1}{2N_{\text{pair}}} \sum_{i=1}^{N_{\text{pair}}} \sum_{c\in\{s,u\}} \|h_{i,c,t_2}^{\ell_m}\|_2}, \tag{5}$$

and the NLL changes of safe and unsafe continuations on $\mathcal{D}_{\text{cal}}$ relative to the unedited model, $\Delta_s^{(m)}(\alpha) = \bar{N}_s^{(m)}(\alpha\delta_m) - \bar{N}_s^{(m)}(0)$ and $\Delta_u^{(m)}(\alpha) = \bar{N}_u^{(m)}(\alpha\delta_m) - \bar{N}_u^{(m)}(0)$. From a pre-specified candidate grid we select the α that maximises $\Delta_u^{(m)}(\alpha) - \Delta_s^{(m)}(\alpha)$ subject to $r_m(\alpha) \le 0.55$, $\Delta_u^{(m)}(\alpha) > 0$, a positive margin between the two classes, and a relative NLL increase on safe continuations of at most 2.5%. The procedure only matches the effective edit scale across models and never sees endpoint test results (Appendix I). Calibration is carried out at the main position t_2, and the position study reuses each model's α_m at every site. Because both the norm of each site-specific direction and the hidden-state norm $\|h_{t_k}^{\ell_m}\|_2$ vary with position, the sweep compares sites at a fixed α_m rather than at matched perturbation norms; norm matching is examined separately in the teacher-forced diagnostic on 8B (Section 5.2).

3.4 Independent linear routing gate

With the edit direction and scale fixed, Stage 3 learns when to use $\alpha_m\delta_m$; the gate does not predict whether a free-running trajectory will be judged Safe or Unsafe. Its data are 900 class-balanced prompts: 450 direct-harmful requests and 450 paired no-op requests, split into 720 for training (360 per class) and 180 for development (90 per class) to set the threshold. Direction training and gate training use different supervision and are never optimised jointly. For each gate prompt the target model natively generates the CoT through t_k and we record the pre-edit $h_{t_k}^{\ell_m}$. With μ_D and μ_N the training means of the direct and no-op representations and c the centre of all training representations,

$$w_g = \frac{\mu_D - \mu_N}{\|\mu_D - \mu_N\|_2}, \qquad s(h) = (h-c)^\top w_g, \qquad g(h) = \mathbf{1}[\, s(h) \ge \tau\,]. \tag{6}$$

The threshold τ is frozen before any test generation: we require a no-op false-trigger rate of at most 5% on the development set, then among feasible thresholds maximise direct recall and then balanced accuracy. $g = 1$ routes to apply and $g = 0$ keeps no-op.

3.5 Conditional single-point intervention

Combining the components frozen in Stages 1–3, Stage 4 follows a strict inference order (Figure 1b): natively generate the early CoT, read the raw state, route, then conditionally edit,

$$\widehat{h}_{t_k}^{\ell_m} = h_{t_k}^{\ell_m} + g\Big(h_{t_k}^{\ell_m}\Big)\; \alpha_m\, \delta_m^{\ell_m}. \tag{7}$$

In the 8B main setting the apply branch adds $\alpha_{8\text{B}}\delta_{8\text{B}}^{20}$ once to the residual state at layer 20, t_2; no later position is edited again, so the edit reaches the remainder of the CoT and the answer only through ordinary autoregressive propagation; the no-op branch follows exactly the computation path of the unedited baseline P0. On 1.5B and 7B the apply branch adds $\alpha_m\delta_m^{17}$ at layer 17, t_2, in the same way, with $\alpha = 8$ and $\alpha = 12$. The model then continues from $\widehat{h}_{t_k}^{\ell_m}$ to generate the rest of the CoT and the final answer. At inference the full method adds one gate evaluation, at most one activation addition, and no base-model parameter update.

3.6 Generation and evaluation implementations

Two evaluation implementations are used. In the three-seed six-arm 8B study, every arm is decoded online. The remaining evaluations use branch composition: for each prompt, P0 and steering share the native prefix through the intervention position t_k and the sampling seed, every token after the edit is regenerated from the edited state, so earlier edits can alter the full downstream trajectory, and the frozen gate then selects one of these two already generated branches, avoiding a third long decode. Position-window claims use the independently generated P0 and steering branches, not branch-composed gated endpoints (Appendix J).

4 EXPERIMENTAL SETUP

Models and data. We evaluate DeepSeek-R1-Distill-Qwen-1.5B/7B and DeepSeek-R1-Distill-Llama-8B with frozen model weights (Guo et al., 2025). Harmful-request evaluation uses StrongREJECT-273, comprising 273 of the 313 StrongREJECT prompts after excluding 40 used during development (Souly et al., 2024), and a 450-prompt SORRY-Bench-derived pool spanning 45 harm categories (Lee et al., 2025; Xie et al., 2025). StrongREJECT is source-independent of direction-training data; the SORRY-derived pool shares its source ecology. Benign-500 contains 500 short, general-purpose Alpaca instructions from the CAST release (Lee et al., 2025; Taori et al., 2023). Normalised exact-text audits find no training–evaluation overlap (Appendix H).

Baselines and evaluation protocol. The main 8B study compares REACT against the unedited model (P0), matched-data CAST and AlphaSteer adaptations, a prompt-boundary gated edit, and a steering-only ablation. All six arms are decoded online on the same 1,223 prompts over three generation seeds. Matched-data baselines use the same target model and supervision resources; these are controlled adaptations, not reproductions of published headline settings. On 1.5B/7B, the corresponding benign evaluation uses one seed and three arms: P0, steering only, and REACT. Separately, three-seed position sweeps cover t_0–t_4 on all three models and both harmful sets, using site-specific directions, ungated edits, and full post-edit regeneration. Positional gains are measured relative to each site's seed-matched P0. Within each model, the sweep fixes the layer and edit scale; perturbation norms can vary across sites.

Judges and metrics. LlamaGuard-3-8B judges the CoT and final answer separately. An output is **Safe** only if structurally valid and both segments are judged safe; valid outputs with either segment unsafe are **Unsafe**, and malformed outputs or missing verdicts are **Invalid**. We report Safe rates, benign refusal, and rubric-scored usefulness (0–100), with mean±sample SD for three-seed results. Separate 8B risk probes are evaluated by AUROC on 272 StrongREJECT and 449 SORRY examples. Full configurations, split construction, and diagnostic protocols are in Appendix A.

5 RESULTS

5.1 THREE-SEED COMPARISON AGAINST MATCHED ALTERNATIVES

Table 1 is the main 8B result. Across three generation seeds, P0 reaches $29.06 \pm 1.29\%$ fully safe CoT–answers on StrongREJECT-273. A matched CAST adaptation improves this to $65.08 \pm 0.85\%$, while the prompt-boundary gated arm reaches $71.31 \pm 3.05\%$. REACT reaches $82.05 \pm 4.03\%$, an increase of 52.99 percentage points over P0 and 10.74 points over the boundary arm in this evaluation. The SORRY-Bench-derived pool shows the same broad pattern: P0 is $47.48 \pm 1.30\%$, CAST $70.00 \pm 1.46\%$, boundary $69.85 \pm 0.64\%$, and REACT $74.96 \pm 0.46\%$. AlphaSteer is close to P0 under this matched-data adaptation. REACT's Invalid rate is 0.00% on StrongREJECT-273 and 0.22% on the SORRY-Bench-derived pool, so the gain is not produced by malformed generations.

The steering-only ablation identifies where the gain comes from. On StrongREJECT-273 it is token-identical to the apply branch selected by REACT, so after hash-aligned score reuse the two harmful-set scores are equal by construction. On the SORRY-Bench-derived pool their Safe rates also differ by less than a point. Thus the direction, not the binary gate, supplies essentially all of the harmful-set improvement in this 8B setting. The gate matters on benign traffic: applying the same edit to every request roughly doubles refusal from 5.33% to 10.67% and lowers usefulness from 91.96 to 87.35, whereas REACT stays at 5.40% refusal and 91.86 usefulness. Across the three seeds the gate fires on only $4.67 \pm 0.42\%$ of Benign-500 requests, and all 1,430 no-op outputs are token-identical to their corresponding P0 outputs. We therefore describe the division of labour precisely: *the direction provides the safety gain; conditional routing prevents paying for it on ordinary benign requests.*

The multi-seed result also clarifies what we do *not* claim. Prompt-boundary intervention is already strong, so the evidence is not that useful control exists only after a reasoning token is generated. Rather, the early-CoT state is a particularly effective one-shot control point under the main matched recipe. An exploratory boundary-initialisation audit in Appendix D shows that boundary performance can move substantially under a different direction initialisation; accordingly, we avoid a universal "t_2 beats prefill" claim.

Table 1: Main 8B six-arm online evaluation over three generation seeds (mean±sample SD). Harmful columns are LlamaGuard fully Safe CoT–answer rates. Benign refusal is lower-is-better; usefulness is on a 0–100 rubric scale.

Method	StrongREJECT-273 Safe ↑	SORRY-derived pool Safe ↑	Benign-500 refusal ↓	Usefulness ↑	Benign trigger
P0	29.06±1.29	47.48±1.30	5.33±0.42	91.96±0.51	–
CAST, matched data	65.08±0.85	70.00±1.46	5.47±0.46	91.80±0.62	1.20±0.00
AlphaSteer, matched data	31.87±0.63	47.63±1.86	5.60±0.69	**92.05**±0.26	N/A
Prompt-boundary gated	71.31±3.05	69.85±0.64	5.47±0.46	91.80±0.51	0.80±0.00
Steering only (no gate)	**82.05**±4.03†	**75.11**±0.38	10.67±0.90	87.35±0.68	100
REACT	**82.05**±4.03†	74.96±0.46	5.40±0.35	91.86±0.45	4.67±0.42

† On StrongREJECT-273, the steering-only ablation and REACT outputs are token-identical and therefore share the same hash-aligned score.

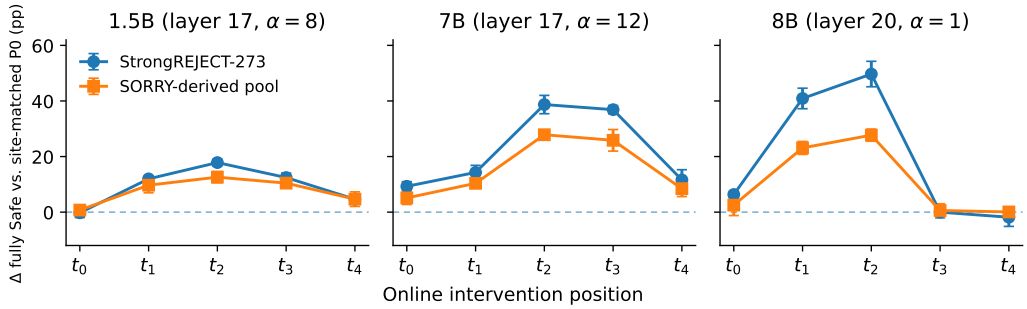

Figure 2: Three-seed position sweeps on all three models with site-specific learned directions and steering-only one-shot edits (8B: layer 20, $\alpha = 1$; 1.5B/7B: layer 17, calibrated α). Points show mean within-seed ΔSafe relative to each site's own P0; error bars are sample SD across three generation seeds. Every post-intervention token is regenerated from the edited state. The early window replicates on both harmful sets and all three models; on 8B it collapses after t_2 (values in Table 2).

5.2 A REPRODUCIBLE EARLY CONTROL WINDOW

The central causal result is replicated across three generation seeds and three models. Figure 2 and Table 2 report the mean within-seed change in fully Safe outputs after one edit, with every token after the intervention position regenerated from the edited state. We use the *steering-only branch* for this localisation so that gate quality cannot create or erase a positional effect. Counting a site as inside the window when its gain reaches at least half of the model's peak, the window is t_1–t_2 on 8B, t_2–t_3 on 7B, and t_1–t_3 on 1.5B, identically on both harmful sets: it is two to three tokens wide and always contains t_2, which is the strongest site on every model and set. On 8B (layer 20) and StrongREJECT-273, the mean ΔSafe is only $+6.35\pm1.12$ points at t_0, then rises to $+40.90\pm3.69$ at t_1 and $+49.69\pm4.58$ at t_2, before collapsing to $+0.00\pm2.04$ at t_3 and -1.83 ± 3.30 at t_4; the SORRY-Bench-derived pool reproduces the same shape ($+2.52\pm3.75$, $+23.11\pm2.32$, $+27.70\pm2.18$, $+0.59\pm2.40$, and $+0.07\pm1.68$ points). On 7B the window covers t_2–t_3 ($+38.71\pm3.30$ and $+36.87\pm1.48$ StrongREJECT points) and on 1.5B t_1–t_3 ($+9.70$ to $+17.83$ across both sets); beyond the window the gain falls to $+11.60\pm3.67$ (7B) and $+4.64\pm1.12$ (1.5B) StrongREJECT points at t_4, whereas 8B returns to its site-matched baseline. Per-seed values, gated endpoints, and trigger rates for all three models are in Appendix D.

The position-specific gates provide a useful control but are not needed for the causal window claim. On 8B their development-set harmful recall is 60.0% at t_0, 100.0% at t_1, 96.7% at t_2, 93.3% at t_3, and only 32.2% at t_4 under a pre-specified benign-FPR$\leq 5\%$ and harmful-recall$\geq 60\%$ qualification rule; on 1.5B and 7B it is 81.1/95.6/93.3/63.3/18.9% and 94.4/96.7/94.4/76.7/38.9% at t_0–t_4, so t_4 fails qualification on every model. Even for the gated endpoints, the 8B t_3 collapse is not a routing artifact: the gate routes 99.6–100% of StrongREJECT and over 90% of SORRY harmful prompts to the steering branch, yet the gated gain stays at baseline (Tables 6–7). We report the gated branch-composition endpoints transparently in Appendix D, but base the temporal conclusion on the independently generated P0 and steering branches above.

Table 2: Three-seed one-shot position sweeps on all three models (Safe %, mean±sample SD; steering-only branch versus site-matched P0). 8B uses layer 20 and $\alpha = 1$; 1.5B and 7B use layer 17 with calibrated $\alpha = 8$ and $\alpha = 12$.

		StrongREJECT-273			SORRY-derived pool		
Model	Site	P0	Steering	Δ (Steering−P0)	P0	Steering	Δ (Steering−P0)
1.5B	t_0	11.36±1.32	11.11±0.56	−0.24±1.69	26.22±0.80	26.96±0.71	+0.74±1.51
	t_1	10.62±0.63	22.59±0.42	+11.97±0.92	26.67±0.80	36.37±2.11	+9.70±2.74
	t_2	9.89±1.60	27.72±1.39	+17.83±1.18	26.30±0.71	38.89±1.82	+12.59±1.99
	t_3	10.38±1.84	22.83±0.92	+12.45±1.60	25.78±0.22	36.22±1.24	+10.44±1.33
	t_4	10.38±2.02	15.02±2.56	+4.64±1.12	26.67±0.80	31.33±1.82	+4.67±2.62
7B	t_0	21.37±1.81	30.65±0.21	+9.28±1.69	39.19±1.00	44.30±3.30	+5.11±2.35
	t_1	21.73±1.29	36.02±2.94	+14.29±2.54	39.56±1.33	49.93±0.64	+10.37±0.84
	t_2	20.15±0.63	58.85±3.88	+38.71±3.30	39.26±0.90	67.11±0.38	+27.85±1.14
	t_3	22.34±1.10	59.22±0.42	+36.87±1.48	39.48±2.53	65.33±3.36	+25.85±3.88
	t_4	21.86±2.02	33.46±1.65	+11.60±3.67	39.33±0.80	47.78±2.04	+8.44±2.84
8B	t_0	28.69±0.92	35.04±1.73	+6.35±1.12	47.70±1.48	50.22±2.31	+2.52±3.75
	t_1	28.94±1.27	69.84±2.77	+40.90±3.69	47.26±1.48	70.37±0.93	+23.11±2.32
	t_2	30.16±2.65	79.85±1.94	+49.69±4.58	46.81±0.46	74.52±2.10	+27.70±2.18
	t_3	29.18±1.88	29.18±0.76	+0.00±2.04	46.96±1.92	47.56±0.67	+0.59±2.40
	t_4	30.65±0.92	28.82±3.05	−1.83±3.30	47.48±1.45	47.56±0.97	+0.07±1.68

Every post-edit token is regenerated; Δ is computed within seed before averaging.

Table 3: Norm-matched local likelihood diagnostic on 8B (276 teacher-forced trajectories, nominal $\ell_2 = 1$). ΔNLL is edited minus P0 NLL for the next ground-truth token; negative is better for that trajectory. Values are trajectory-level mean ± sample SD.

Site	$D_{\mathrm{KL}}(P_0 \| P_{\mathrm{edit}})$	Safe ΔNLL	Unsafe ΔNLL
t_0	0.000002±0.000004	−0.000009±0.000017	−0.000010±0.000019
t_1	0.01347±0.00930	+0.00925±0.14065	+0.00705±0.12153
t_2	**0.09273**±0.05110	**−0.21314**±0.58805	**+0.10061**±0.44777
t_3	0.02488±0.02357	+0.20097±0.29933	+0.31064±0.31645
t_4	0.00095±0.00236	−0.01213±0.05653	−0.00368±0.02811

Local likelihood structure of the window. On 8B, a norm-matched teacher-forced diagnostic gives a more specific account than perturbation magnitude alone. Table 3 applies each learned direction at nominal ℓ_2 norm 1 to the same 276 trajectories and measures the immediate output-distribution KL and the next-target-token NLL change. The t_2 edit is the only strong site with the desired class-selective sign pattern: it makes the next safe target more likely ($\Delta\text{NLL} = -0.213$) while making the next unsafe target less likely ($+0.101$). By contrast, t_3 substantially changes the output distribution but raises NLL for both safe and unsafe targets ($+0.201 / +0.311$), indicating strong but poorly aligned local influence rather than insufficient edit magnitude. At t_2, the learned-direction KL (0.0927) is about $196\times$ the mean KL of three fixed norm-matched random directions (0.000473). The same sign distinction between t_2 and t_3 persists at nominal norms 0.5, 1, and 2 (Appendix D).

This diagnostic is deliberately limited. In fixed-continuation measurements the large t_2 KL contracts sharply after the first future token and is no longer dominant at longer offsets; we therefore do *not* claim that t_2 works because its perturbation simply remains larger throughout the trajectory. The evidence instead supports a narrower mechanism correlate: under the learned direction family, successful endpoint steering coincides with a locally selective safe-versus-unsafe likelihood shift that is absent at the failed 8B t_3 site.

5.3 Readability and causal controllability diverge

If the early peak simply reflected risk information becoming more legible as generation proceeds, later states should be both easier to classify and easier to steer. The readout experiment gives the opposite pattern. On 8B and StrongREJECT-273, a frozen prompt-only risk probe reaches AUROC 0.648; reading t_2 increases this to 0.714, and reading t_4 increases it again to 0.750. On the SORRY-Bench-derived pool, prompt-only AUROC is 0.659 and the strongest prompt-plus-CoT readout reaches

Table 4: Online risk readout on 8B without steering. The frozen probability ensemble is evaluated once on 272 StrongREJECT and 449 SORRY examples; these are probe AUROCs, not averages over three generation seeds. Higher late-state readability does not imply stronger one-shot controllability.

Probe Input	AUROC	
	StrongREJECT	SORRY
Prompt only	0.648	0.659
t_2 only	0.714	0.676
t_4 only	**0.750**	0.675
Prompt + t_2	0.722	0.692
Prompt + t_4	0.747	**0.699**

0.699. Thus late states can remain at least as predictive of risk even when one-shot steering has already lost its causal effect.

This distinction changes the role of the probe analysis. We no longer use readability to select t_2. Instead, the combined evidence says that *predictive information and causal leverage follow different trajectories*: t_4 is a better StrongREJECT risk readout than t_2, but the 8B t_4 intervention is ineffective. The control site must therefore be established causally, not selected from probe AUROC alone.

5.4 Robustness across scale and depth

In the three-seed sweeps, the one-shot t_2 edit raises StrongREJECT Safe from $9.89 \pm 1.60\%$ to $27.72 \pm 1.39\%$ on 1.5B and from $20.15 \pm 0.63\%$ to $58.85 \pm 3.88\%$ on 7B, and the SORRY-derived pool from $26.30 \pm 0.71\%$ to $38.89 \pm 1.82\%$ and from $39.26 \pm 0.90\%$ to $67.11 \pm 0.38\%$ (Table 2). The single-seed cross-model comparison corroborates these endpoints within three points under gating and endpoint-label-free strength calibration, supplies the only 1.5B/7B Benign-500 endpoints, and cross-checks LlamaGuard with an independent rubric evaluator that moves in the same direction on every model and set (full Safe/Unsafe/Invalid decompositions in Appendix F). On Benign-500, gating returns refusal to the P0 level on both models (7.20% and 9.00%) and keeps usefulness within 0.05 points of P0, whereas removing the gate raises refusal to 14.60% and 15.80%. At 8B, a single-seed layer sweep found effective t_2 edits at layers 9, 17, 20, 25 and 32, with StrongREJECT Safe rates of 76.92–80.95% under gated site-specific steering (Appendix G). These results show that the core phenomenon is not tied to one hidden size or one transformer depth, while the post-edit-regeneration position sweeps show that it is highly sensitive to *when* in the generated reasoning trajectory the edit is made.

6 Conclusion

REACT identifies an early window for one-shot safety steering within the first five CoT tokens. Three-seed position sweeps with site-specific directions and full post-edit regeneration show that, on 8B, a layer-20 edit is weak at t_0, strongly effective at t_1/t_2, and returns to approximately the site-matched baseline at t_3/t_4 on both harmful sets. The same sweep on 1.5B and 7B reproduces the early window with boundaries shifted by at most one token, while cross-layer results further support the t_2 effect.

Risk readability follows a different pattern: on 8B and StrongREJECT, the fifth-token state predicts risk better than the third-token state despite offering little safety gain under the evaluated edit. A norm-matched local diagnostic also distinguishes useful control from raw sensitivity: t_2 increases the next safe target's likelihood and decreases the unsafe target's, whereas t_3 suppresses both. These results support selecting intervention sites through their effects on generation, rather than probe performance alone, without implying universal superiority over prefill.

The main three-seed, six-arm 8B evaluation demonstrates the practical benefit: the t_2 REACT intervention raises complete-trajectory safety by roughly 53 percentage points on source-independent StrongREJECT and 27 points on the SORRY-derived pool, outperforming the evaluated matched alternatives while keeping benign refusal and usefulness near P0. The edit supplies the harmful-set gain; conditional routing limits its cost on benign requests.

Reproducibility Statement

Section 3 and Appendix A specify the intervention, routing, generation, and evaluation contracts. Appendix H documents split construction and exact-overlap audits; Appendix J records generation details; Appendix K records judge configuration; and Appendix L describes the artifacts planned for release. Per-seed endpoint summaries and the local likelihood diagnostics used in the control-window analysis are reported in Appendix D.

Ethics Statement

This work studies a defensive intervention intended to reduce unsafe reasoning and answers. We evaluate on existing harmful-request benchmarks and report aggregate scores rather than reproducing harmful generations. The method can in principle be reversed or miscalibrated to weaken safeguards, and its gate may shift under domain or population change; these limitations and deployment considerations are discussed in Appendix N. No human-subject study was conducted.

AI Use Statement

Generative AI tools were used to assist with language editing, organization, LaTeX formatting, and presentation of author-provided experimental results. They were not used to generate the reported experimental measurements. The authors reviewed the resulting text, tables, and numerical claims and take responsibility for the submission.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4322.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

Shourya Batra, Pierce Tillman, Samarth Gaggar, Shashank Kesineni, Kevin Zhu, Sunishchal Dev, Ashwinee Panda, Vasu Sharma, and Maheep Chaudhary. SALT: Steering activations towards leakage-free thinking in chain of thought. *arXiv preprint arXiv:2511.07772*, 2025.

Soumya Suvra Ghosal, Souradip Chakraborty, Vaibhav Singh, Furong Huang, Dinesh Manocha, and Amrit Singh Bedi. Safety recovery in reasoning models is only a few early steering steps away. *arXiv preprint arXiv:2602.11096*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*, 2024.

Melody Y. Guan, Miles Wang, Micah Carroll, Zehao Dou, Annie Y. Wei, Marcus Williams, Benjamin Arnav, Joost Huizinga, Ian Kivlichan, Mia Glaese, Jakub Pachocki, and Bowen Baker. Monitoring monitorability. *arXiv preprint arXiv:2512.18311*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Peixuan Han, Cheng Qian, Xiusi Chen, Yuji Zhang, Heng Ji, and Denghui Zhang. SafeSwitch: Steering unsafe LLM behavior via internal activation signals. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 37, 2024. doi: 10.52202/079017-0261.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama Guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. SafeChain: Safety of language models with long chain-of-thought reasoning capabilities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23303–23320, 2025.

Ishita Kakkar, Enze Zhang, Rheeya Uppaal, and Junjie Hu. HarmThoughts: A benchmark for fine-grained harmful behavior detection in reasoning traces. *arXiv preprint arXiv:2604.19001*, 2026. Accepted at COLM 2026.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, et al. Chain of thought monitorability: A new and fragile opportunity for AI safety. *arXiv preprint arXiv:2507.11473*, 2025.

Tharindu Kumarage, Ninareh Mehrabi, Anil Ramakrishna, Xinyan Zhao, Richard Zemel, Kai-Wei Chang, Aram Galstyan, Rahul Gupta, and Charith Peris. Towards safety reasoning in LLMs: AI-agentic deliberation for policy-embedded CoT data creation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22694–22715, 2025. doi: 10.18653/v1/2025.findings-acl.1166.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *International Conference on Learning Representations*, 2025.

Changyi Li, Jiayi Wang, Xudong Pan, Geng Hong, and Min Yang. ReasoningShield: Content safety detection over reasoning traces of large reasoning models. *arXiv preprint arXiv:2505.17244*, 2025.

Xiaomin Li, Jianheng Hou, Zheyuan Deng, Zhiwei Zhang, Taoran Li, Binghang Lu, Bing Hu, Yunhan Zhao, and Yuexing Hao. Chain of risk: Safety failures in large reasoning models and mitigation via adaptive multi-principle steering. *arXiv preprint arXiv:2605.05678*, 2026.

Mansi Phute, Alec Helbling, Matthew Hull, ShengYun Peng, Sebastian Szyller, Cory Cornelius, and Duen Horng Chau. LLM self defense: By self examination, LLMs know they are being tricked. *arXiv preprint arXiv:2308.07308*, 2024. Tiny Papers Track, ICLR 2024.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828.

Alexander Robey, Eric Wong, Hamed Hassani, and George J. Pappas. SmoothLLM: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*, 2023.

Leheng Sheng, Changshuo Shen, Weixiang Zhao, Junfeng Fang, Xiaohao Liu, Zhenkai Liang, Xiang Wang, An Zhang, and Tat-Seng Chua. AlphaSteer: Learning refusal steering with principled null-space constraint. In *International Conference on Learning Representations*, 2026.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for empty jailbreaks. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 37, 2024. doi: 10.52202/079017-3984.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Raj Vardhan Tomar, Preslav Nakov, and Yuxia Wang. UnsafeChain: Enhancing reasoning model safety via hard cases. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2025*, pages 1233–1247, 2025. doi: 10.18653/v1/2025.findings-ijcnlp.75.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jiayou Wang, Rundong Liu, Yue Hu, Huijia Wu, and Zhaofeng He. SecDecoding: Steerable decoding for safer LLM generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20504–20521, 2025a. doi: 10.18653/v1/2025.findings-emnlp.1118.

Yuquan Wang, Mi Zhang, Yining Wang, Geng Hong, Mi Wen, Xiaoyu You, and Min Yang. ReasoningGuard: Safeguarding large reasoning models with inference-time safety aha moments. *arXiv preprint arXiv:2508.04204*, 2025b.

Zijun Wang, Haoqin Tu, Yuhan Wang, Juncheng Wu, Yanqing Liu, Jieru Mei, Brian R. Bartoldson, Bhavya Kailkhura, and Cihang Xie. STAR-1: Safer alignment of reasoning LLMs with 1k data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 37988–37997, 2026. doi: 10.1609/aaai.v40i44.41136.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Xiaofei Wen, Wenxuan Zhou, Wenjie Jacky Mo, and Muhao Chen. ThinkGuard: Deliberative slow thinking leads to cautious guardrails. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13698–13713, 2025. doi: 10.18653/v1/2025.findings-acl.704.

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

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. SORRY-bench: Systematically evaluating large language model safety refusal. In *International Conference on Learning Representations*, 2025.

Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. Defending ChatGPT against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5(12):1486–1496, 2023.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. 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, 2024. doi: 10.18653/v1/2024.acl-long.303.

Kia-Jüng Yang, Dominik Meier, Jiachen Zhao, Terry Ruas, and Bela Gipp. Beyond a single direction: Chain-of-thought disrupts simple steering of refusal. *arXiv preprint arXiv:2605.26772*, 2026.

Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. On prompt-driven safeguarding for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 61593–61613, 2024. URL https://proceedings.mlr.press/v235/zheng24n.html.

A EXPERIMENTAL SETUP

Research questions. **RQ1**: does one early native-CoT edit improve full-trajectory safety relative to the unedited model and matched activation-steering alternatives? **RQ2**: where is a one-shot edit causally effective when every post-edit token is regenerated from the edited state, and does this coincide with where risk is easiest to read? **RQ3**: what local probability changes distinguish effective from ineffective intervention sites under matched perturbation norms? **RQ4**: can conditional routing retain the safety gain without the benign cost of applying the same edit to every request, and does the effect replicate across model sizes and transformer depths?

Models and intervention configuration. We evaluate DeepSeek-R1-Distill-Qwen-1.5B, DeepSeek-R1-Distill-Qwen-7B, and DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025), always with frozen base-model parameters. Each model uses one configuration across all of its experiments: 8B edits layer 20 with $\alpha = 1$, and 1.5B and 7B edit layer 17 with endpoint-label-free strengths $\alpha = 8$ and $\alpha = 12$ calibrated at t_2. The main configuration intervenes at t_2; the position study retrains a site-specific direction at each of $t_0, \ldots, t_4$ on every model with the same layer and α and a site-specific gate.

Evaluation sets. StrongREJECT-273 (StrongREJECT-273) is the source-independent harmful generalisation test: it contains the 273 of the 313 official StrongREJECT prompts that remain after removing 40 used during method development, by deterministic normalised exact-text matching (Souly et al., 2024). The SORRY-Bench-derived pool contains 450 prompts over 45 harm categories from the public CAST/SORRY-Bench-v1 ecology (Lee et al., 2025; Xie et al., 2025); because the direction-training prompts come from the same source ecology, we treat this set as within-ecology rather than source-independent evidence. Benign-500 contains 500 Alpaca benign test instructions from the CAST release (Lee et al., 2025; Taori et al., 2023) and is used for false-trigger, refusal, and usefulness auditing. Direction learning uses the 170 exact-prompt safe/unsafe pairs of Section 3.1.

Main six-arm online comparison. The main 8B experiment contains six arms: P0 (no intervention); full REACT (the t_2 gate plus edit); matched-data CAST and AlphaSteer adaptations; a prompt-boundary gated edit; and a *steering-only* t_2 ablation that removes the gate and applies the same edit to every request. "Matched-data" means the baseline mechanism is fitted to the same target model and frozen supervision/evaluation ecology, rather than reproducing published headline settings. Each arm is generated online for three seeds on all 1,223 prompts ($22{,}014$ prompt–seed–arm units); we report mean and sample SD. On StrongREJECT-273, REACT and steering-only outputs are token-identical, so identical hashes reuse one canonical judge score.

Cross-model comparison. On 1.5B and 7B we compare three arms—P0, steering-only, and full REACT—on all 1,223 prompts with a single seed, at t_2 and layer 17 with each model's calibrated α. This unit provides the only Benign-500 endpoints for these two models and cross-checks LlamaGuard with an independent rubric evaluator (Appendix F).

Three-seed causal position sweep. To compare intervention timing, each site t_k generates through t_k natively, applies its site-specific edit once, and regenerates every later token from the edited state. Thus an early edit can change the full downstream trajectory. We repeat the sweep on all three models and both harmful sets for three generation seeds; each model/site/seed has its own P0, so position claims use within-seed steering-minus-P0 differences. Gated endpoints in Appendix D only select between these already generated branches and are routing diagnostics, not evidence for the temporal window.

Norm-matched local likelihood diagnostic. To probe why the learned directions differ across positions, we teacher-force the same 276 offline trajectories (138 exact-prompt safe/unsafe pairs) through the 8B model and rescale each site-specific learned direction to nominal ℓ_2 norm 1. At the intervention logits we measure $D_{\mathrm{KL}}(P_0 \| P_{\mathrm{edit}})$ and the change in NLL of the next ground-truth token, separately for safe and unsafe trajectories. The 32 pairs used for signed direction optimisation are excluded from this diagnostic, although the 138-pair pool participated in mean-difference initialisation. Three fixed unit random directions provide a norm-matched reference. Norm sweeps and fixed-

continuation downstream KL measurements appear in Appendix D. This teacher-forced analysis is diagnostic rather than a free-running causal explanation.

Risk readout diagnostic. Separately from steering, on 8B we evaluate frozen probability-ensemble probes that read the prompt state, t_2, t_4, or concatenated prompt-plus-CoT features and predict trajectory risk on online outputs. These are readout diagnostics, not routing labels for the main method. The StrongREJECT and SORRY readout sets contain 272 and 449 evaluable examples respectively.

Judges and metrics. LlamaGuard-3-8B (Inan et al., 2023; Grattafiori et al., 2024) judges the complete CoT and answer separately. A sample is **Safe** only if it is structurally valid and both segments are safe, **Unsafe** if valid with either segment unsafe, and **Invalid** if it violates the output-structure contract or a judge verdict is missing or cannot be parsed; malformed output therefore cannot masquerade as a safety repair. Benign refusal and adapted usefulness (0–100) are scored with the rubric evaluator described in Appendix K. Unless explicitly stated otherwise, percentage-point differences are descriptive; three-seed tables report mean$\pm$sample-standard-deviation rather than inferential confidence intervals.

B LIMITATIONS

Window boundaries differ across models. The t_0–t_4 position sweep, which regenerates every post-edit token, uses three independent generation seeds, reproduces on two harmful sets, and covers all three models; the early window replicates, but its boundaries shift by up to one token, and model size, target layer (17 versus 20), and edit strength co-vary across these runs. The cross-size sweeps establish that the early window replicates, not that the exact t_1/t_2 boundaries observed on 8B are universal. Within each model the sweep holds α_m fixed rather than matching perturbation norms across sites; on 8B, the norm-matched diagnostic shows that the t_2/t_3 contrast persists at nominal norms 0.5, 1, and 2.

The local likelihood analysis is a mechanism correlate, not a complete causal explanation. The norm-matched diagnostic is teacher-forced, computed on 8B, and measures immediate next-token KL/NLL changes on a fixed trajectory corpus. It shows that on 8B, t_2 has a selectively aligned safe/unsafe local effect while t_3 does not, but fixed-continuation downstream KL does not remain uniquely large for t_2. We therefore do not claim to have identified the dynamical mechanism that makes the free-running t_2 edit effective.

Boundary performance depends on the direction-learning recipe. The main prompt-boundary baseline is a matched-data gated adaptation under the primary recipe. In a separate single-seed diagnostic, changing the boundary direction initialisation while retaining signed optimisation substantially changes endpoint safety (Appendix D). We therefore claim an effective early-CoT control point under our evaluated recipe, not universal superiority of t_2 to every possible prefill objective.

Readout and routing are not the same task. The online risk probes predict trajectory risk, whereas the deployed REACT gate is trained on direct-harmful versus paired no-op prompts. A controlled single-seed comparison on 8B that holds the t_2 edit fixed finds similar harmful-set endpoints for prompt-state and t_2 gates (Appendix D). We therefore do not claim that the gate is a fine-grained trajectory-risk detector; its demonstrated role is primarily to avoid unnecessary benign edits.

Baseline adaptations are controlled comparisons. CAST and AlphaSteer are fitted to the 8B target model and frozen data/evaluation ecology. This improves experimental comparability but is not a reproduction of every training choice or benchmark setting in the original publications; the comparison supports performance under our contract rather than a global ranking of the methods.

Source ecology and overlap. StrongREJECT-273 is source-independent of the direction-training material, whereas the SORRY-Bench-derived pool shares the CAST/SORRY-Bench ecology with it. A normalized exact-prompt audit using Unicode NFKC normalisation, whitespace collapsing, and case folding finds no overlap between either the 170 direction prompts or the 900 gate prompts and any endpoint evaluation set. The direction and gate training resources share 88 prompts. This establishes exact-text separation from endpoints, not semantic disjointness or independence of the shared source ecology; a full near-duplicate audit is not available.

Benign evaluation is not a capability benchmark. Benign-500 shows that aggregate ordinary-instruction refusal and usefulness remain near P0, but it is not a hard-benign, mathematics, coding, knowledge, or complex-reasoning benchmark. We therefore make no broad capability-preservation claim.

Automatic judges and scope. Safety is measured with LlamaGuard and rubric-based automatic evaluators. All evaluated base models belong to the DeepSeek-R1 distilled family, spanning two base architectures (Qwen2.5 and Llama-3.1). Conclusions are limited to these models, datasets, steering objectives, and generation contracts.

C EXPANDED 8B ONLINE EVALUATION

The main six-arm study contains three online generation seeds for each of 1,223 prompts and six arms, for 22,014 prompt–seed–arm output units. The main text reports mean±sample-standard-deviation across seeds. Safe denotes LlamaGuard safety of the complete CoT–answer and is distinct from refusal or usefulness.

Table 5: Expanded 8B online results. The StrongREJECT steering-only and REACT outputs are token-identical; the canonical table therefore reuses the same score for identical hashes.

Method	StrongREJECT-273 Safe	SORRY-derived pool Safe	Benign-500 refusal	Usefulness
P0	29.06 ± 1.29	47.48 ± 1.30	5.33 ± 0.42	91.96 ± 0.51
CAST, matched data	65.08 ± 0.85	70.00 ± 1.46	5.47 ± 0.46	91.80 ± 0.62
AlphaSteer, matched data	31.87 ± 0.63	47.63 ± 1.86	5.60 ± 0.69	92.05 ± 0.26
Prompt-boundary gated	71.31 ± 3.05	69.85 ± 0.64	5.47 ± 0.46	91.80 ± 0.51
Steering only (no gate)	82.05 ± 4.03	75.11 ± 0.38	10.67 ± 0.90	87.35 ± 0.68
REACT	82.05 ± 4.03	74.96 ± 0.46	5.40 ± 0.35	91.86 ± 0.45

Across the three seeds REACT's benign trigger rate is $4.67 \pm 0.42\%$, versus 1.20% for the CAST adaptation, 0.80% for the prompt-boundary gate, and 100% for the steering-only ablation. The 1,430 REACT no-op benign outputs are token-identical to P0. REACT Invalid rates are 0.00% on StrongREJECT and 0.22% on the SORRY-derived pool.

D POSITION SWEEP AND STEERING DIAGNOSTICS

Three-seed position endpoints. For each site and seed, the model generates through the intervention site natively, applies the site-specific learned direction once, and regenerates every subsequent token from the edited state. P0 and steering are therefore independently generated continuations after the intervention point. Because each site has its own P0, the relevant statistic is the within-seed difference. Table 2 in Section 5.2 reports means and sample SDs across three generation seeds for all three models.

Per-seed endpoint values. Tables 6–7 expose the per-seed 8B percentages underlying Table 2, and Tables 8–11 give the same for 1.5B and 7B. "Gated" is a branch-composition diagnostic: the frozen site-specific gate reads the natural prefix and selects the already generated P0 or steering continuation; it is not a third online long-form generation.

Cross-model position sweeps. The same three-seed, five-site sweep is repeated on 1.5B and 7B with each model's frozen configuration (layer 17; calibrated $\alpha = 8$ and $\alpha = 12$), site-specific learned directions, and site-specific gates fitted under the same 720/180 protocol. The t_2 gate is the same frozen gate as in the single-seed cross-model comparison, and the other four sites' gates are fitted separately; each gate is frozen once per model and site, and the three seeds are generation seeds, not gate refits. Tables 8–11 use the notation of Table 6.

Gate qualification. Each site-specific harmful/no-op gate is fitted on 720 prompts (360/360) and thresholded on a 180-prompt development set (90/90). Threshold selection maximises harmful recall subject to benign FPR$\leq 5\%$; a gate is considered qualified only if harmful recall is at least 60%. Table 12 gives the 8B statistics and Table 13 the 1.5B/7B statistics: on every model t_0–t_3 qualify,

Table 6: 8B StrongREJECT-273 per-seed position results (%). Trigger is the fraction routed to the steering branch. †: the t_4 gate fails the pre-specified development qualification rule.

Site	Seed	P0 Safe	Steering Safe	Gated Safe	Gated Unsafe	Gated Invalid	Trigger
t_0	20260831	28.57	33.70	32.97	67.03	0.00	72.89
	20260832	27.84	34.43	31.87	68.13	0.00	72.89
	20260833	29.67	37.00	33.70	65.93	0.37	72.89
t_1	20260831	27.47	71.79	71.79	28.21	0.00	100.00
	20260832	29.67	66.67	66.67	33.33	0.00	100.00
	20260833	29.67	71.06	71.06	28.94	0.00	100.00
t_2	20260831	31.50	79.12	79.12	20.88	0.00	100.00
	20260832	31.87	78.39	78.39	21.61	0.00	100.00
	20260833	27.11	82.05	82.05	17.95	0.00	100.00
t_3	20260831	29.67	30.04	30.04	69.96	0.00	99.63
	20260832	27.11	28.94	28.94	70.70	0.37	99.63
	20260833	30.77	28.57	28.57	71.43	0.00	100.00
$t_4^\dagger$	20260831	31.50	26.37	30.04	69.96	0.00	60.81
	20260832	29.67	27.84	29.67	70.33	0.00	57.51
	20260833	30.77	32.23	30.40	69.60	0.00	58.61

Table 7: 8B SORRY-derived pool per-seed position results (%). Notation as in Table 6.

Site	Seed	P0 Safe	Steering Safe	Gated Safe	Gated Unsafe	Gated Invalid	Trigger
t_0	20260831	47.33	51.56	48.22	51.11	0.67	50.22
	20260832	49.33	47.56	48.44	51.11	0.44	50.67
	20260833	46.44	51.56	48.67	51.11	0.22	50.89
t_1	20260831	46.00	70.67	70.67	29.11	0.22	97.78
	20260832	48.89	69.33	69.33	30.67	0.00	97.78
	20260833	46.89	71.11	71.11	28.22	0.67	97.78
t_2	20260831	46.67	76.89	76.89	22.89	0.22	98.89
	20260832	47.33	73.78	73.78	26.22	0.00	98.44
	20260833	46.44	72.89	72.89	26.67	0.44	98.67
t_3	20260831	47.33	46.89	46.89	52.67	0.44	92.00
	20260832	48.67	47.56	47.33	52.67	0.00	90.44
	20260833	44.89	48.22	48.67	51.33	0.00	90.67
$t_4^\dagger$	20260831	48.89	48.22	47.78	51.56	0.67	31.11
	20260832	47.56	46.44	48.22	51.78	0.00	31.78
	20260833	46.00	48.00	46.89	52.22	0.89	30.22

whereas t_4 does not. The 8B position-sweep gates are distinct from those of the single-seed layer ablation unit (Appendix G), which were fitted separately; the 8B t_0 gate here meets the 60% recall threshold exactly.

Norm-matched local likelihood diagnostic. The mechanism diagnostic uses all 276 trajectories from 138 exact-prompt safe/unsafe pairs in the non-optimisation pool: 138 Safe and 138 Unsafe trajectories, with no missing positions. These 138 prompts do not overlap the separate 32-pair signed-optimisation subset, although all 170 pairs participated in the initial mean-difference direction. Every position uses the same 276 trajectories. Computation is teacher-forced with fixed future text and no gate. We define $D_{\mathrm{KL}}(P_0 \| P_{\mathrm{edit}})$ at the intervention logits and ΔNLL as edited minus P0 NLL for the next ground-truth token. Table 14 reports trajectory-level mean$\pm$sample SD at nominal $\ell_2 = 1$.

Three fixed unit random vectors (seeds 202609140–202609142) are reused across every position and trajectory. Their between-direction mean$\pm$sample SD is shown in Table 15. At t_2, the learned-direction mean KL is approximately $196\times$ the random-direction mean. Actual BF16 additive norms remain within 0.25% of nominal norm 1 for both learned and random directions, with zero failures under the pre-specified 10% tolerance.

Table 8: 1.5B StrongREJECT-273 per-seed position results (%), layer 17, $\alpha = 8$. Notation as in Table 6.

Site	Seed	P0 Safe	Steering Safe	Gated Safe	Gated Unsafe	Gated Invalid	Trigger
t_0	20260831	12.82	10.62	12.45	76.19	11.36	70.70
	20260832	10.26	10.99	10.99	77.29	11.72	75.09
	20260833	10.99	11.72	12.09	75.82	12.09	70.70
t_1	20260831	11.36	22.34	22.34	71.43	6.23	95.97
	20260832	10.26	22.34	22.34	71.06	6.59	95.60
	20260833	10.26	23.08	22.71	71.43	5.86	95.60
t_2	20260831	10.99	29.30	28.94	70.33	0.73	98.17
	20260832	8.06	26.74	25.64	70.33	4.03	97.80
	20260833	10.62	27.11	27.47	68.50	4.03	95.97
t_3	20260831	12.09	22.71	23.44	71.79	4.76	91.94
	20260832	8.42	21.98	22.34	71.79	5.86	93.41
	20260833	10.62	23.81	23.81	73.26	2.93	89.38
$t_4^\dagger$	20260831	11.72	17.58	13.92	77.66	8.42	36.63
	20260832	8.06	12.45	10.99	80.22	8.79	36.26
	20260833	11.36	15.02	15.38	75.09	9.52	33.33

Table 9: 1.5B SORRY-derived pool per-seed position results (%). Notation as in Table 6.

Site	Seed	P0 Safe	Steering Safe	Gated Safe	Gated Unsafe	Gated Invalid	Trigger
t_0	20260831	26.44	26.67	26.44	60.67	12.89	67.56
	20260832	26.89	26.44	27.11	63.78	9.11	66.44
	20260833	25.33	27.78	27.11	60.89	12.00	67.33
t_1	20260831	26.44	38.44	37.33	54.44	8.22	88.00
	20260832	27.56	34.22	34.44	58.22	7.33	88.44
	20260833	26.00	36.44	36.00	56.44	7.56	87.33
t_2	20260831	27.11	38.44	37.33	56.44	6.22	90.44
	20260832	25.78	37.33	36.22	56.00	7.78	88.44
	20260833	26.00	40.89	39.78	53.33	6.89	90.67
t_3	20260831	26.00	36.44	35.11	58.00	6.89	63.11
	20260832	25.56	37.33	34.67	60.00	5.33	63.78
	20260833	25.78	34.89	33.56	59.78	6.67	63.56
$t_4^\dagger$	20260831	26.89	30.89	26.89	63.33	9.78	16.67
	20260832	27.33	29.78	26.67	59.78	13.56	18.22
	20260833	25.78	33.33	25.56	62.22	12.22	15.56

Norm sweep. Table 16 shows the immediate diagnostic at nominal norms 0.5, 1, and 2. The qualitative t_2 sign pattern (safe ΔNLL negative, unsafe positive) persists throughout; t_3 instead raises both class NLLs at every tested norm. This rules out a simple "late edit is too weak" account for the t_3 failure under this direction family.

Fixed-continuation downstream KL. To test whether t_2 simply creates a larger perturbation throughout the future computation, we keep the original future text fixed and re-read output distributions at offsets $q \in \{0, 1, 2, 4, 8, 16, 32, 64\}$. Table 17 shows that the large immediate t_2 KL rapidly contracts; by $q = 2$ and later it is not consistently larger than the other sites. This negative result is why we frame the immediate NLL pattern as a mechanism correlate rather than a complete propagation mechanism.

Prompt-state gate versus t_2-state gate. A single-seed isolation control holds the t_2 direction and edit location fixed and changes only where the harmful/no-op gate is read. The harmful endpoints are nearly identical, while the prompt gate triggers less often on benign inputs. This supports the narrower claim that routing protects benign traffic rather than a claim that the t_2 gate is itself a uniquely powerful trajectory-risk detector.

Boundary-initialisation sensitivity. A separate single-seed audit shows that prompt-boundary steering is sensitive to how the direction optimisation is initialised. Both rows below use signed

Table 10: 7B StrongREJECT-273 per-seed position results (%), layer 17, $\alpha = 12$. Notation as in Table 6.

Site	Seed	P0 Safe	Steering Safe	Gated Safe	Gated Unsafe	Gated Invalid	Trigger
t_0	20260831	20.51	30.77	30.77	48.72	20.51	91.21
	20260832	23.44	30.77	30.77	46.89	22.34	90.84
	20260833	20.15	30.40	28.57	50.92	20.51	91.21
t_1	20260831	23.08	38.83	38.83	60.07	1.10	98.90
	20260832	21.61	32.97	32.60	65.20	2.20	98.90
	20260833	20.51	36.26	36.63	61.90	1.47	98.90
t_2	20260831	19.78	58.24	57.88	41.76	0.37	98.53
	20260832	19.78	55.31	55.31	44.69	0.00	98.53
	20260833	20.88	63.00	63.00	37.00	0.00	97.44
t_3	20260831	21.25	59.71	59.34	40.66	0.00	95.24
	20260832	23.44	58.97	59.34	40.29	0.37	95.24
	20260833	22.34	58.97	58.61	40.66	0.73	93.41
$t_4^\dagger$	20260831	23.81	31.87	27.11	69.96	2.93	63.37
	20260832	21.98	33.33	27.11	70.33	2.56	60.81
	20260833	19.78	35.16	28.94	69.96	1.10	62.64

Table 11: 7B SORRY-derived pool per-seed position results (%). Notation as in Table 6.

Site	Seed	P0 Safe	Steering Safe	Gated Safe	Gated Unsafe	Gated Invalid	Trigger
t_0	20260831	39.11	45.11	44.89	40.22	14.89	81.78
	20260832	38.22	40.67	42.44	41.11	16.44	81.78
	20260833	40.22	47.11	45.33	40.67	14.00	83.11
t_1	20260831	39.56	49.56	49.33	48.44	2.22	94.22
	20260832	38.22	49.56	49.33	49.11	1.56	94.67
	20260833	40.89	50.67	50.44	48.44	1.11	94.44
t_2	20260831	38.22	67.33	66.44	33.33	0.22	88.67
	20260832	39.78	66.67	65.56	33.78	0.67	87.33
	20260833	39.78	67.33	66.44	32.89	0.67	88.67
t_3	20260831	36.67	65.78	62.22	36.67	1.11	77.56
	20260832	40.22	61.78	60.89	38.44	0.67	78.67
	20260833	41.56	68.44	66.89	32.67	0.44	76.44
$t_4^\dagger$	20260831	39.11	48.22	41.78	56.67	1.56	37.78
	20260832	40.22	45.56	42.00	57.11	0.89	37.11
	20260833	38.67	49.56	41.78	56.89	1.33	36.89

Table 12: 8B site-specific gate development statistics.

Site	AUROC	Harmful recall	Benign FPR	Threshold	Qualified
t_0	0.9563	60.00%	2.22%	1.0618	yes
t_1	0.9984	100.00%	4.44%	-1.1934	yes
t_2	0.9967	96.67%	4.44%	-1.1597	yes
t_3	0.9781	93.33%	4.44%	-0.0849	yes
t_4	0.7988	32.22%	1.11%	9.4828	no

continuation optimisation; this is *not* a mean-difference-only versus optimised-direction ablation. This sensitivity is why the paper does not claim that an early-CoT site intrinsically dominates every prompt-boundary objective.

Alternative routers and lookahead. Using the same t_2 steering direction in a strict online single-seed unit, the deployed harmful/no-op gate reaches 78.02%/75.11% Safe on StrongREJECT/SORRY from a common 27.84%/48.00% P0, while an aligned t_2 risk-probe router reaches 75.82%/72.67%. In a separate common-P0 unit, prompt-boundary gated steering reaches 74.36%/69.11%, direct t_2 REACT 76.92%/75.33%, and a t_4 risk probe followed by rollback to t_2 steering 73.63%/73.11%. The benign trigger rates are 0.80%, 4.80%, and 9.20% respectively. On StrongREJECT, the paired 95%

Table 13: 1.5B/7B site-specific gate development statistics under the same protocol as Table 12.

Model	Site	Harmful recall	Benign FPR	Qualified
1.5B	t_0	81.11%	4.44%	yes
	t_1	95.56%	2.22%	yes
	t_2	93.33%	4.44%	yes
	t_3	63.33%	3.33%	yes
	t_4	18.89%	1.11%	no
7B	t_0	94.44%	3.33%	yes
	t_1	96.67%	4.44%	yes
	t_2	94.44%	4.44%	yes
	t_3	76.67%	3.33%	yes
	t_4	38.89%	4.44%	no

Table 14: Learned-direction immediate diagnostic at nominal $\ell_2 = 1$. KL uses 276 trajectories; Safe/Unsafe ΔNLL use 138 trajectories each.

Site	KL mean±SD	Safe ΔNLL mean±SD	Unsafe ΔNLL mean±SD
t_0	$2.392\times10^{-6}\pm3.958\times10^{-6}$	$-9.308\times10^{-6}\pm1.668\times10^{-5}$	$-9.538\times10^{-6}\pm1.907\times10^{-5}$
t_1	0.013468 ± 0.009297	$+0.009251\pm0.140650$	$+0.007049\pm0.121532$
t_2	0.092733 ± 0.051099	-0.213144 ± 0.588052	$+0.100614\pm0.447770$
t_3	0.024877 ± 0.023566	$+0.200971\pm0.299331$	$+0.310636\pm0.316452$
t_4	0.000952 ± 0.002364	-0.012130 ± 0.056530	-0.003684 ± 0.028106

Table 15: Norm-matched random-direction baseline: mean±sample SD across three fixed random directions.

Site	KL	Safe ΔNLL	Unsafe ΔNLL
t_0	$(3.883\pm1.852)\times10^{-7}$	$(2.659\pm1.230)\times10^{-6}$	$(2.091\pm1.153)\times10^{-6}$
t_1	0.0003378 ± 0.0000274	$+0.0008448\pm0.0014854$	$+0.0016015\pm0.0008751$
t_2	0.0004732 ± 0.0000937	$+0.0033526\pm0.0110436$	$+0.0036887\pm0.0010349$
t_3	0.0009461 ± 0.0001540	-0.0010160 ± 0.0028581	-0.0093058 ± 0.0044889
t_4	0.0003032 ± 0.0000481	-0.0017813 ± 0.0027533	-0.0005377 ± 0.0006652

Table 16: Immediate learned-direction diagnostic across nominal edit norms.

Site	Norm	Mean KL	Mean Safe ΔNLL	Mean Unsafe ΔNLL
t_0	0.5	7.324×10^{-7}	-4.825×10^{-6}	-5.278×10^{-6}
	1	2.392×10^{-6}	-9.308×10^{-6}	-9.538×10^{-6}
	2	8.120×10^{-6}	-1.539×10^{-5}	-1.541×10^{-5}
t_1	0.5	0.003373	+0.000609	+0.001028
	1	0.013468	+0.009251	+0.007049
	2	0.056962	+0.047224	+0.027941
t_2	0.5	0.018899	-0.125694	+0.021434
	1	0.092733	-0.213144	+0.100614
	2	0.586086	-0.118908	+0.605321
t_3	0.5	0.006243	+0.089256	+0.136033
	1	0.024877	+0.200971	+0.310636
	2	0.114529	+0.503777	+0.783656
t_4	0.5	0.000314	-0.006735	-0.002590
	1	0.000952	-0.012130	-0.003684
	2	0.003301	-0.018147	-0.005954

interval for the t_2-minus-boundary difference in that unit is $[-4.03, +8.79]$ points. These diagnostics motivate no claim that the deployed gate is a fine-grained trajectory-risk predictor; its demonstrated role is primarily to avoid unnecessary benign edits.

Table 17: Mean $D_{\mathrm{KL}}(P_0 \| P_{\mathrm{edit}})$ at downstream offsets for the learned direction, nominal $\ell_2 = 1$.

Site	$q = 0$	$q = 1$	$q = 2$	$q = 4$	$q = 8$	$q = 16$	$q = 32$	$q = 64$
t_0	0.000002392	0.000286935	0.000523217	0.000188030	0.000202658	0.000211524	0.000365771	0.000492711
t_1	0.013468359	0.011323466	0.000733684	0.000385064	0.000286980	0.000189030	0.000269300	0.000465836
t_2	0.092732667	0.001113817	0.000175178	0.000184779	0.000196407	0.000229168	0.000339613	0.000392153
t_3	0.024876751	0.000185371	0.000349253	0.000280625	0.000183828	0.000220134	0.000319331	0.000427258
t_4	0.000951569	0.000542396	0.000153110	0.000184515	0.000150855	0.000184716	0.000345119	0.000371083

Table 18: Single-seed gate-observation isolation with the same layer-20 t_2 edit ($\alpha = 1$).

Method	StrongREJECT-273 Safe	SORRY-derived pool Safe	Benign-500 trigger
Prompt-state gate $\rightarrow t_2$ edit	78.75	76.44	0.80
t_2-state gate $\rightarrow t_2$ edit	79.12	76.89	4.60

Table 19: Prompt-boundary direction-initialisation audit.

Boundary direction training	StrongREJECT-273 P0→Safe	SORRY-derived pool P0→Safe
Mean-difference init. + signed optimisation	28.94→59.34	46.89→62.22
Zero init. + signed optimisation	28.94→78.02	46.89→71.56

E ADDITIONAL TEACHER-FORCED PROBE PROTOCOL AND FULL RESULTS

This appendix documents the auxiliary probing analysis of the main readability analysis. Everything here is correlational and teacher-forced; nothing in it enters the intervention pipeline, the gate, or the strength calibration.

Corpus. The probe corpus shares no source dataset with any evaluation set in the paper (it draws on none of the CAST-derived, SORRY-Bench-derived, StrongREJECT, or Alpaca material); this is source-level disjointness by construction, and no string-level overlap audit was run. It pools nine raw sources with trajectory-level safe/unsafe labels: the ReasoningShield training (SFT and DPO) and test splits (Li et al., 2025) (mixed-label), two STAR safe-reasoning sources (Wang et al., 2026), two AIDSAFE policy-embedded safe-reasoning sources (Kumarage et al., 2025), UnsafeChain corrected safe trajectories (Tomar et al., 2025), and an unsafe-reasoning source drawn from HarmThoughts (Kakkar et al., 2026). Partial, garbage, and unlabelled trajectories are dropped; each source is capped at 1,000 examples; prompts, reasoning, and final answers are capped at 800, 1,600, and 800 words. The ReasoningShield test split is held out from training as a whole and provides the *held-out* metric; *test* denotes the in-distribution test split of the pooled training sources. For leave-one-source-out (LOSO) evaluation the sources are grouped into six label-balanced families: the three ReasoningShield splits, and three mixed-source families in which the safe-only STAR, AIDSAFE, and UnsafeChain sources are each paired with the HarmThoughts unsafe source (because a safe-only source has no defined AUROC on its own).

Protocol. Each trajectory is teacher-forced through the frozen target model (DeepSeek-R1-Distill-Qwen-1.5B or DeepSeek-R1-Distill-Llama-8B) and hidden states are read at fixed positions: `last_prompt_token` (strict prompt-only; final user-prompt token before the assistant turn), `pre_think` (immediately before `<think>`), `think_last` (the `<think>` boundary token; the strongest pre-CoT control, not a strict prompt-only baseline), and the zero-based natural CoT tokens `cot_2`, `cot_3`, `cot_4`, `cot_7`, where `cot_0` is the first CoT token; `cot_2` is therefore the intervention site t_2. Only linear probes are used, one per (position, layer), over a dense layer grid; the scan configuration enumerates 16 layers for 1.5B (from layer 4 to 28) and 19 layers for 8B (from layer 4 to 32). Following the source protocol, the construct-validity comparison uses the across-layer mean (and median) rather than the best layer, so that the result does not rest on a best-of-scan choice. Positions t_0 and t_1 are not part of this control scan.

Reading the tables. Prompt-only and pre-CoT states already carry substantial signal: on the in-distribution test split the strict prompt-only mean is 0.931 (1.5B) and 0.919 (8B), and the best

Table 20: Across-layer AUROC of linear probes, 1.5B. "Test" is the in-distribution test split; "held-out" is the held-out ReasoningShield test source. The paper reports held-out mean.

Position	Test best	Test mean	Test median	Held-out best	Held-out mean	Held-out median
`last_prompt_token`	0.952	0.931	0.932	0.832	0.785	0.787
`pre_think`	0.955	0.945	0.944	0.855	0.830	0.834
`think_last`	0.954	0.939	0.938	0.846	0.833	0.836
`cot_2` (t_2)	0.966	0.954	0.953	0.895	0.867	0.875
`cot_3` (t_3)	0.979	0.971	0.970	0.902	0.868	0.873
`cot_4`	0.978	0.972	0.973	0.928	0.891	0.894
`cot_7`	0.976	0.963	0.961	0.888	0.826	0.843

Table 21: Across-layer AUROC of linear probes, 8B. Columns as in Table 20.

Position	Test best	Test mean	Test median	Held-out best	Held-out mean	Held-out median
`last_prompt_token`	0.945	0.919	0.922	0.831	0.801	0.802
`pre_think`	0.958	0.939	0.939	0.868	0.843	0.848
`think_last`	0.954	0.939	0.941	0.881	0.859	0.870
`cot_2` (t_2)	0.967	0.956	0.954	0.911	0.886	0.893
`cot_3` (t_3)	0.985	0.980	0.980	0.933	0.912	0.917
`cot_4`	0.982	0.976	0.976	0.946	0.921	0.929
`cot_7`	0.977	0.971	0.972	0.910	0.862	0.861

Table 22: Held-out mean AUROC of linear probes over the layer grid (separate probe corpus; teacher-forced, correlational). t_2 is the intervention site; the `<think>` boundary token is the strongest pre-CoT control.

Position	1.5B	8B
Last prompt token (strict prompt-only)	0.785	0.801
`<think>` boundary token (strongest pre-CoT control)	0.833	0.859
t_2 (third CoT token; intervention site)	0.867	0.886
t_3 (fourth CoT token)	0.868	0.912

prompt-only layer reaches 0.952 and 0.945. The held-out gap is therefore the relevant comparison, and at t_2 it is $+0.082$/$+0.085$ over strict prompt-only but only $+0.034$/$+0.027$ over `think_last`. The readout keeps increasing at t_3 and t_4, so a ranking by readout would favour later positions than the one we intervene at; the intervention site is chosen by the causal ablation of Section 5.2, not by this table.

Leave-one-source-out by family. Table 23 lists the best position-scan AUROC when each family is held out. Two mixed-source families, `aidsafe` and `unsafechain`, are almost perfectly separable already at the first or second CoT token (`cot_0` or `cot_1`; 1.000 and 0.995 for both models), which the source analysis attributes to source or format artifacts; they are diagnostic splits and not trajectory-monitoring evidence. The macro means over all six families (1.5B 0.926 ± 0.071; 8B 0.927 ± 0.064) are inflated by those two families and should not be quoted as a transfer result. On the three ReasoningShield families the 8B trajectory-minus-prompt gap is only $+0.001$, $+0.013$, and $+0.066$.

Same-prompt diagnostic. A stricter control comes from natural same-prompt pair corpora, in which a safe and an unsafe continuation are sampled from the same prompt so that the prompt carries no label information by construction; the resolvable provenance of these prompts is ReasoningShield-only. On the 8B-generated corpus (467 pairs; 374/47/46 train/validation/test), a prompt-only TF-IDF classifier scores test AUROC 0.5000, the best prompt or pre-CoT hidden-state probe with the 8B extractor scores 0.5000, and the validation-selected early-CoT hidden probe scores 0.6983; on the same corpus, length alone reaches 0.8261 and word-level TF-IDF over the CoT text 0.9660. On a 32B-generated corpus (359 pairs), the prompt or pre-CoT control is 0.5000 with the 8B extractor and about 0.51 with a 32B extractor, while validation-selected early-CoT probes reach 0.7188 (8B extractor) and 0.7890 (32B extractor); a generated-safe/original-unsafe variant (663 pairs) shows

Table 23: LOSO family-level AUROC (best position/layer in the position scan) and the trajectory-minus-prompt gap reported by the source analysis. Rows marked † are artifact-prone diagnostic splits; RS = ReasoningShield.

	1.5B			8B		
Held-out family	Best site	AUROC	Gap	Best site	AUROC	Gap
RS train (SFT split)	cot_3/L16	0.908	+0.062	cot_3/L18	0.909	+0.001
RS train (DPO split)	cot_1/concat	0.787	−0.006	cot_3/concat	0.807	+0.013
RS test	cot_4/L17	0.932	+0.079	cot_3/L14	0.920	+0.066
STAR vs. HarmThoughts	cot_4/L17	0.933	+0.015	cot_4/L12	0.929	+0.012
AIDSAFE vs. HarmThoughts†	cot_0/L4	1.000	+0.120	cot_0/L4	1.000	+0.118
UnsafeChain vs. HarmThoughts†	cot_1/L7	0.995	+0.083	cot_0/L6	0.995	+0.086

length-only 0.8817 and word TF-IDF 0.9423. Surface baselines are therefore strong in every variant, and the hidden-state signal is not shown to be cleaner than surface features.

What this analysis does not prove. It does not show that h_{t_2} predicts the outcome of an online rollout, that editing h_{t_2} changes the outcome (that is the intervention experiment), that the probed signal is safety semantics rather than trajectory or format regularities, or that t_2 is the best site among early tokens. It supports one statement: on a separate corpus, an above-chance linearly readable signal beyond the prompt is present at t_2.

F Cross-model results: rubric evaluator and LlamaGuard three-arm decomposition

Table 24 gives the single-seed independent rubric-evaluator results for all three arms of every model; Section 5.4 reports the 1.5B and 7B gated values in the text. The single-seed 8B rows are separate from the three-seed six-arm comparison.

Table 24: Cross-model results under the independent rubric evaluator (%): P0, steering only (no gate), and gated REACT with model-specific α_m and gates.

Model (α_m)	Arm	StrongREJECT-273 refusal	SORRY-derived pool refusal	Benign-500 refusal	Benign-500 usefulness
1.5B (8)	P0	17.58	52.00	7.20	79.30
	Steering only	44.69	55.78	14.60	73.80
	Gated	**43.59**	**55.78**	**7.20**	**79.25**
7B (12)	P0	31.14	48.22	9.00	87.23
	Steering only	73.26	63.78	15.80	81.35
	Gated	**72.53**	**62.89**	**9.00**	**87.23**
8B (1)	P0	45.42	50.89	4.20	93.98
	Steering only	85.35	64.44	10.60	87.35
	Gated	**85.35**	**64.44**	**4.00**	**94.18**

Table 25 gives the exactly paired Safe/Unsafe/Invalid decomposition for all three models, matching each prompt, its first three natural CoT tokens, and its sampling seed across the three arms. Gated steering raises the Safe share on both harmful sets without shifting Unsafe mainly into Invalid; relative to steering-only intervention it gives up a little harmful-set gain and returns Benign-500 to the neighbourhood of P0. The 1.5B/7B matrix covers 7,338 prompt–arm units; the 8B rows come from a single-seed exact-pair unit (1,223 prompts × 3 arms), not the three-seed six-arm comparison.

G Cross-layer robustness at t_2

The single-seed layer sweep is a robustness check for the t_2 site. It is a separate single-seed generation unit with site-specific directions and gates at fixed $\alpha = 1$. We retain only the cross-layer t_2 comparisons here; position rows from a design that fixed future prefix tokens are not used for any temporal claim, which instead relies on the post-edit-regeneration sweep in Appendix D. The corresponding independent rubric evaluation gives StrongREJECT refusal/risk of 86.45/12.09

Table 25: Single-seed LlamaGuard Safe / Unsafe / Invalid (%) under exact pairing for the 1.5B ($\alpha = 8$), 7B ($\alpha = 12$), and 8B ($\alpha = 1$) models.

		P0			Steering only			Gated		
Model	Dataset	S	U	I	S	U	I	S	U	I
1.5B	StrongREJECT-273	8.42	75.82	15.75	28.57	68.13	3.30	**27.47**	**69.23**	**3.30**
1.5B	SORRY-derived pool	27.78	62.22	10.00	42.44	50.44	7.11	**41.33**	**51.33**	**7.33**
1.5B	Benign-500	92.40	0.00	7.60	96.80	0.00	3.20	**92.60**	**0.00**	**7.40**
7B	StrongREJECT-273	21.98	76.19	1.83	59.71	40.29	0.00	**59.34**	**40.66**	**0.00**
7B	SORRY-derived pool	40.44	57.78	1.78	68.00	31.78	0.22	**67.11**	**32.22**	**0.67**
7B	Benign-500	97.60	0.00	2.40	98.40	0.00	1.60	**97.60**	**0.00**	**2.40**
8B	StrongREJECT-273	27.11	72.53	0.37	79.49	20.15	0.37	**79.49**	**20.15**	**0.37**
8B	SORRY-derived pool	49.56	50.00	0.44	73.56	26.44	0.00	**73.56**	**26.44**	**0.00**
8B	Benign-500	98.80	0.00	1.20	99.40	0.00	0.60	**99.00**	**0.00**	**1.00**

Table 26: 8B cross-layer t_2 robustness (LlamaGuard fully Safe fraction, %).

Site	StrongREJECT-273 steering only	StrongREJECT-273 gated	SORRY-derived pool steering only	SORRY-derived pool gated	Benign-500 trigger
P0	26.74		49.11		–
L9–t_2	82.05	80.95	75.56	73.33	1.60
L17–t_2	77.66	77.66	74.22	74.44	2.20
L20–t_2	76.92	76.92	74.00	74.00	4.80
L25–t_2	78.75	78.75	72.22	72.22	1.00
L32–t_2	79.12	79.12	74.67	74.67	3.40

(L9), 87.55/11.45 (L17), 86.81/11.63 (L20), 86.45/12.18 (L25), and 87.91/10.35 (L32), versus P0 45.05/48.08. Benign refusal remains 5.0–5.2% and usefulness 92.83–92.98 across these t_2 sites.

H Data construction and split hygiene

Table 27: Evaluation sets and their roles. Only StrongREJECT-273 is source-independent of the direction-training material.

Dataset	Size	Role	Relation to direction-training data
StrongREJECT-273 (Souly et al., 2024)	273	main harmful generalisation test	source-independent
SORRY-Bench-derived pool (Lee et al., 2025; Xie et al., 2025) (SORRY-derived pool in tables)	450	45-category full-trajectory safety	same CAST/SORRY-Bench ecology
Benign-500 (Lee et al., 2025; Taori et al., 2023)	500	false triggers, over-refusal, usefulness	CAST Alpaca benign test split

StrongREJECT-273. The official StrongREJECT benchmark has 313 prompts. Forty prompts were used during method development; the frozen evaluation subset excludes exactly those 40 by normalised exact-text matching and keeps all remaining 273. The construction script fixes the input hash, the exclusion list, and the output-set hash, and will be released with the code so that the subset can be regenerated deterministically.

SORRY-Bench-derived pool. The 450-prompt, 45-category set is the SORRY-Bench-v1-derived pool contained in the public CAST release (Lee et al., 2025), built on the June 2024, 45-category taxonomy of SORRY-Bench (Xie et al., 2025). It is not the official SORRY-Bench base set (440 instructions over 44 categories in the published version), and we never report it as such.

Benign-500. Benign-500 is not a separately released dataset: it is the `test` array of `docs/demo-data/alpaca.json` in the public CAST release (Lee et al., 2025), i.e., the 500 held-out harmless prompts of CAST's own evaluation, whose instructions come from the Stanford Alpaca instruction-following data (Taori et al., 2023). It contains 500 unique single-turn English

instructions (4–19 words, median 9) covering general-purpose tasks; the most frequent leading verbs are *Generate*, *Create*, *Name*, *Write*, *Explain*, *Describe*, and *Rewrite*. The release stores only the instruction text, so the 173 items (34.6%) whose Alpaca source entry has a separate input field (e.g., "Rewrite the following sentence…") are presented without that input. Every arm receives identical prompts, so this affects absolute refusal and usefulness levels but not between-arm comparisons. The split is disjoint from the 700-instruction training split of the same file and from the harmful direction-training data, and we use it only for false-trigger, refusal, and usefulness auditing.

Direction-training pairs. The 170 exact-prompt pairs are frozen offline trajectories whose prompts come from the CAST release (direct requests and CAST-derived variants on the SORRY-Bench-v1 taxonomy); construction, re-encoding, and the 32-pair optimisation subset are described in Section 3.1. Because of this shared ecology, results on the SORRY-Bench-derived pool are reported as within-ecology evidence and StrongREJECT-273 carries the source-independent claim.

Gate prompts and normalized exact-overlap audit. The gate uses 900 class-balanced prompts (450 direct-harmful, 450 paired no-op) split 720/180 into training and development. The same prompt sets and split are used for all three models; only the target model's own t_2 representations differ. We re-audit raw prompt fields using Unicode NFKC normalisation, whitespace collapsing, and case folding. Table 28 shows zero normalized exact overlap from either training resource into any endpoint evaluation set. The direction and gate resources share 88 prompts, all in the gate training split, so the two training components are not mutually prompt-disjoint. StrongREJECT-273 and the SORRY-derived pool share two normalized exact prompts with each other; this is evaluation-set overlap, not train–test overlap. The audit establishes normalized exact-text separation from training to endpoints, not semantic or near-duplicate disjointness.

Table 28: Normalized exact-prompt overlap after Unicode NFKC, whitespace collapsing, and case folding.

	Gate-900	StrongREJECT-273	SORRY-derived pool	Benign-500
Direction-170	88	0	0	0
Gate-900	–	0	0	0
StrongREJECT-273	–	–	2	0
SORRY-derived pool	–	–	–	0

Gate train and gate development have zero normalized exact overlap; all 88 direction–gate overlaps lie in gate train.

I Strength-calibration protocol

After the direction is frozen, the calibration pool $\mathcal{D}_{\text{cal}}$ draws a fixed 128 safe and 128 unsafe trajectories from the 138 pairs that did not enter the signed-likelihood optimisation. For the 1.5B and 7B models, the intervention layer is fixed to layer 17 before calibration. For each candidate α on a pre-specified grid, we run teacher-forced forward passes with $h_{t_2} \leftarrow h_{t_2} + \alpha\delta$ and record the mean continuation NLL of each class and the two NLL changes Δ_s, Δ_u of Section 3.3 on $\mathcal{D}_{\text{cal}}$, together with the relative perturbation $r_m(\alpha)$, whose denominator is the mean t_2 hidden-state norm over all 340 initialisation trajectories. A candidate is feasible if $r_m(\alpha) \leq 0.55$, $\Delta_u > 0$, the margin $\Delta_u - \Delta_s$ is positive, and the relative NLL increase on safe continuations is at most 2.5%. Among feasible candidates we select the largest margin. The selected strengths are $\alpha = 12$ for 7B and $\alpha = 8$ for 1.5B, both at the pre-fixed layer 17; the 8B main setting instead inherits the pre-fixed reference configuration at layer 20 and $\alpha = 1$, which was not selected by this calibration procedure. The calibration never reads StrongREJECT-273, SORRY-Bench-derived pool, or Benign-500 outputs or labels.

J Generation contracts

All endpoint experiments use temperature 0.6, top-p 0.95, and a budget of 4,096 new tokens for the continuation; in the true-online runs, tokens generated before the intervention position do not count

against this budget, so a complete output can exceed 4,096 tokens by the length of that prefix. The **main 8B six-arm study** samples every arm online for three seed indices. REACT first generates through t_2 natively, reads the gate state, and then either continues from the edited t_2 state or follows the no-op computation. The **three-seed position study** runs on all three models (8B at layer 20 with $\alpha = 1$; 1.5B/7B at layer 17 with $\alpha = 8$ and $\alpha = 12$) and independently generates a site-matched P0 continuation and a steering-only continuation at each of $t_0, \ldots, t_4$; after the intervention, every later token is resampled from the edited state and no future token is fixed. The site-specific gated rows in Appendix D are branch compositions: the frozen gate reads the natural prefix and selects between those already generated P0 and steering continuations, so the gated rows are routing diagnostics rather than additional long-form generations. The position-sweep executions use seed indices 20260831–20260833 on every model. An execution-level generation batch adjustment from 64 to 32 occurred within the 8B sweep, so we treat the three runs as independent generation replications rather than controlled pairs that differ only by RNG seed. Model, layer, α, decoding hyperparameters, site-specific directions, and endpoint judge are otherwise fixed. The **single-seed 1.5B/7B cross-model comparison** (t_2 only) also uses paired branch composition: P0 and steering begin from the same native t_0, t_1, t_2 prefix and sampling seed, the steering continuation is generated after the t_2 edit, and the frozen gate selects one of the two resulting continuations. This avoids a third long decode for the gated arm. It is used for cross-size robustness, not for the temporal position-window claim. Outputs without an explicit closed CoT followed by a final answer are counted Invalid rather than repaired or resampled.

K JUDGE CONFIGURATION

LlamaGuard evaluation uses Llama-Guard-3-8B (Grattafiori et al., 2024). The CoT and the final answer of each output are judged as two separate segments; Safe requires structural validity and a safe verdict on both. Independent evaluation uses the benchmark authors' released rubrics: the StrongREJECT rubric yields refusal and rubric risk scores, while the SORRY-Bench score-only rubric yields compliance, with refusal defined as $1 -$ compliance. For Benign-500 we report refusal and adapted usefulness from the StrongREJECT rubric. The judge model depends on the experimental unit, as specified below. We persist only parsed scores, model identities, and prompt/response hashes, not raw judge completions or protected text; identical prompt/response pairs reuse their existing score rather than being re-scored.

Judge versions and fallbacks. For the three-seed, six-arm 8B study, API calls request `gpt-4o-mini-2024-07-18`, and all completed API scores return that same model version. Benign coverage is complete for all three seeds; all arms use the same rubric family and identical prompt–response hashes reuse one parsed score.

The following judge versions and fallback counts apply only to the single-seed cross-model and ablation evaluations, not to the three-seed six-arm study. These calls requested `gpt-4o-mini` (StrongREJECT rubric and Benign-500; official `strongreject_rubric` template, commit `7a551d5b`) or `gpt-4o` (SORRY-Bench judge); their primary returned versions were `gpt-4o-mini-2024-07-18` and `gpt-4o-2024-08-06`. On these single-seed harmful sets, a small fraction of source-arm response-score cells was returned by a fallback model and is retained as scored: StrongREJECT-273 14/546 cells for 1.5B (12 `gpt-4o-2024-08-06`, 2 `gpt-4.1-mini-2025-04-14`), 5/546 for 7B (3 and 2), and 2/546 for the single-seed 8B unit (`gpt-4o-2024-08-06`); SORRY-derived pool 18/900 for 1.5B and 20/900 for 7B (`gpt-4o-mini-2024-07-18`), and 0/900 for the single-seed 8B unit; 8B ablation, StrongREJECT-273 0/3,822 and SORRY-derived pool 595/6,300 cells (588 unique prompt–response pairs; `gpt-4o-mini-2024-07-18`). These denominators count the P0/steering source-arm cells; gated score reuse does not constitute additional judge calls.

L ARTIFACTS AND REPRODUCIBILITY

Upon publication we will release the code, the frozen steering directions and gates for all three models, the generation contract, the deterministic construction script for the StrongREJECT-273 subset, per-seed position-sweep summaries, local-diagnostic measurements, and the per-prompt routing logs that the audits in Appendix M refer to.

M AUDIT COUNTS

The main 8B comparison covers 22,014 online prompt–seed–arm units: 1,223 prompts $\times$ six arms $\times$ three seeds. Benign rubric coverage is complete for 9,000 prompt–seed–arm cells across the six arms. For REACT, 1,430 benign no-op outputs are token-identical to their corresponding P0 outputs. On StrongREJECT, REACT and the steering-only ablation are token-identical, so the canonical analysis reuses one score per identical prompt–response hash.

Each model's three-seed position sweep contains $723 \times 5 \times 3 = 10{,}845$ site–prompt–seed units per source branch, hence 21,690 independently generated P0/steering long-form outputs per model and 65,070 across the three models. Gated position endpoints reuse these outputs by branch composition and do not add long-form generations. The local likelihood diagnostic processes all 276 trajectories at all five positions; all 35 expected execution shards completed and passed the supplied completeness checks, with no trajectory excluded for missing positions.

The 1.5B/7B exact-pair matrix covers 7,338 prompt–arm units (1,223 prompts $\times$ three arms $\times$ two models). The cross-layer 8B robustness and the three-seed t_0–t_4 position sweeps are separate frozen generation units and are never pooled with the main six-arm endpoints. This separation is intentional: each table reports comparisons only within its own generation contract.

N BROADER IMPACT

REACT is a defensive mechanism intended to reduce unsafe reasoning and answers without broadly withdrawing useful access. The main 8B steering-only ablation quantifies why conditional routing matters: refusal on ordinary benign instructions rises from 5.33% to 10.67% and usefulness falls from 91.96 to 87.35, while gated REACT remains near P0. We have not evaluated real deployment traffic or hard-benign populations, so this aggregate benign result should not be treated as a deployment guarantee. As with other steering techniques, reversing or miscalibrating an intervention could weaken safeguards, and automatic judges can mislabel.

Distribution shift and gate calibration. The linear gate is trained on English direct-harmful versus paired no-op prompts and its threshold is set on that development distribution. Domain, language, dialect, or user-population shift can change its trigger distribution. Under-triggering removes protection; over-triggering reintroduces the refusal cost of applying the edit indiscriminately. Any deployment should therefore recalibrate and monitor the gate on target-domain benign traffic rather than assume the reported threshold transfers unchanged.